Integrating Business Intelligence and Machine Learning for Enhanced Decision Making and Data Handling in Real Estate

Arsalan Ameer Khan (21175562)

Supervisor: Dr. Parnia Samimi

18th September 2023

A dissertation to be submitted in partial fulfilment of the requirements for the degree of Master of Science in Big Data Analytics School of Computing, Engineering and the Built Environment Birmingham City University

School of Computing, Engineering and the Built Environment
Birmingham City University

Abstract

This study intends to investigate how Business intelligence and Machine learning are applied to real estate data in the Canadian market. It will examine the many strategies, techniques, and instruments used in this situation and point out their advantages and disadvantages. Additionally, the paper will include a thorough analysis of pertinent studies, research papers, and business reports that have advanced understanding in this field. Also, this study aims to shed light on the potential of BI and ML in empowering decision makers in the Canadian real estate market by evaluating the existing literature and gaining insights from real-world applications. The importance of data-driven decision making in a sector where precise forecasting, risk management, and market information are crucial will be emphasised. The project demonstrates how Raw data is extracted from the website and carefully select and stored in PostgreSQL which was then used to manage data and transform it through a pipeline in Pentaho. Making various reports in Power BI later for improved decision-making. Additionally, a proper feature selection is carried out by correlation analysis prior to the machine learning algorithm's implementation. Random oversampling is also used to address the issue of an unbalanced dataset. To improve the model's accuracy, this is being done. To choose the best suitable algorithm, the model's accuracy has been evaluated in comparison to other algorithms, such as Random Forest, Linear Regression, Elastic.net and much more.

Keywords: Business Intelligence, Machine learning, PostgreSQL, Pentaho, PowerBI, Machine Learning Algorithms.

Acknowledgements

I would like to express my sincere gratitude to all those who have supported and guided me throughout my journey in completing this Master's Report. Your contributions, encouragement, and unwavering belief in my abilities have been instrumental in making this research a reality.

Also, I extend my heartfelt thanks to my family for their unwavering support and understanding throughout this academic endeavour. Your encouragement and belief in my aspirations have been a constant source of motivation.

Table of Contents

List of Figures

Chapter 1: Introduction

Data is a priceless resource in the current digital era that is transforming sectors all over the world, and the real estate industry is no different. The incorporation of cutting-edge data handling techniques is undergoing a significant transition in the traditionally rooted real estate sector. In the past, real estate decisions were frequently made purely on gut feeling and scant knowledge. However, the development of technology has brought to an era of abundant data. Stakeholders now have access to a wide variety of data, including transaction histories, demographic insights, and market trends as well as property listings. The purchase, sale, management, and valuation of real estate are being redefined by this data-driven revolution. A variety of data sources, such as real estate firms, property portals, public documents, and user-generated material, are used by the real estate sector to provide a vast amount of information. For the purpose of creating a complete and accurate dataset for analysis, these many data streams must be gathered, combined, and cleaned. Managing real estate data becomes more difficult as both its volume and variety increase. For decision-making to be effective, data needs to be organised, stored, and made available. Strong data management systems are required to handle massive amounts of data while maintaining their integrity and security. The way real estate professionals approach decision-making and data management has changed dramatically as a result of the marriage of business intelligence (BI) and machine learning (ML). This article examines the integration of BI and ML in the context of real estate projects, illuminating how this synergy helps stakeholders to make knowledgeable decisions and successfully negotiate the market's intricacies. An important aspect of Canada's economic environment is the real estate market. It is one of the biggest sectors, makes up a sizeable portion of the GDP, and is a crucial gauge of the state of the economy as a whole. For investors, developers, legislators, and other stakeholders, the real estate industry in Canada is dynamic, which brings both opportunities and difficulties. (Smith J. , 2019)

Reliable information and efficient analysis are essential for making judgements in this complex economy. Traditional data analysis techniques frequently fall short of capturing the complex relationships and patterns found in the real estate market. However, the development of business intelligence (BI) and machine learning (ML) approaches has created new opportunities for enhancing decision-making by leveraging the power of real estate data.

Business intelligence encompasses a range of technologies and methodologies that enable organizations to collect, analyse, and interpret large volumes of data to gain valuable insights. (Smith J. , 2019) ML algorithms, on the other hand, facilitate the automated learning and extraction of patterns from data, thereby enabling predictive modelling and more accurate decision making. Data usage in real estate goes beyond simple record-keeping. Predictive modelling and machine learning algorithms are two examples of analytical methods that can be used to analyse market patterns, project property values, and even estimate future demand. Stakeholders are better equipped to make informed decisions and foresee changes in the market thanks to these insights. (Thompson, 2020)

1.1 Aims and Objectives

- **Aims:**

Aims for this project is to investigate the possibility for combining machine learning and business intelligence methods in the real estate industry and to investigate how decision-making might be improved in the real estate market by using business intelligence and machine learning. Also, to create cutting-edge data handling techniques for real estate data using BI and machine learning.

- **Objectives:**

1. Identify the main issues that the real estate business faces with decision-making and data processing.
2. Examine how business intelligence contributes to the process of drawing conclusions from real estate data.
3. Look into several machine learning techniques that are appropriate for analysing and predicting real estate data.
4. Create a framework for incorporating business intelligence and machine learning strategies into the choice-making process for real estate.
5. Utilising dataset related to real estate, implement and assess the proposed framework.
6. Analyse the effect of the integrated strategy on the quality and effectiveness of decision-making in the real estate industry.
7. Give advice on how business intelligence and machine learning solutions should be used and used in the real estate sector.

Chapter 2: Background and Rationale

In (Dumitru-Alexandru, 2016) the advent of Big Data has facilitated the rise of Business Intelligence (BI) and its application in the field of economics. The authors of this reference explored the utilization of financial derivatives systems by investment banks to manage closed funds with remarkably high return ratios of 250:1. Building upon this, they further developed a Business-Automated Data Economy Model, drawing inspiration from IBM's Cognos Workflow, to enable automated decision making in the public sector, such as in education and healthcare.

In a similar vein, (Scholly, 2019) delved into the integration of Big Data within BI and its potential impact on decision-making processes within the realm of public housing. The authors proposed a comprehensive framework and architecture to support and enhance decision-making activities in this domain.

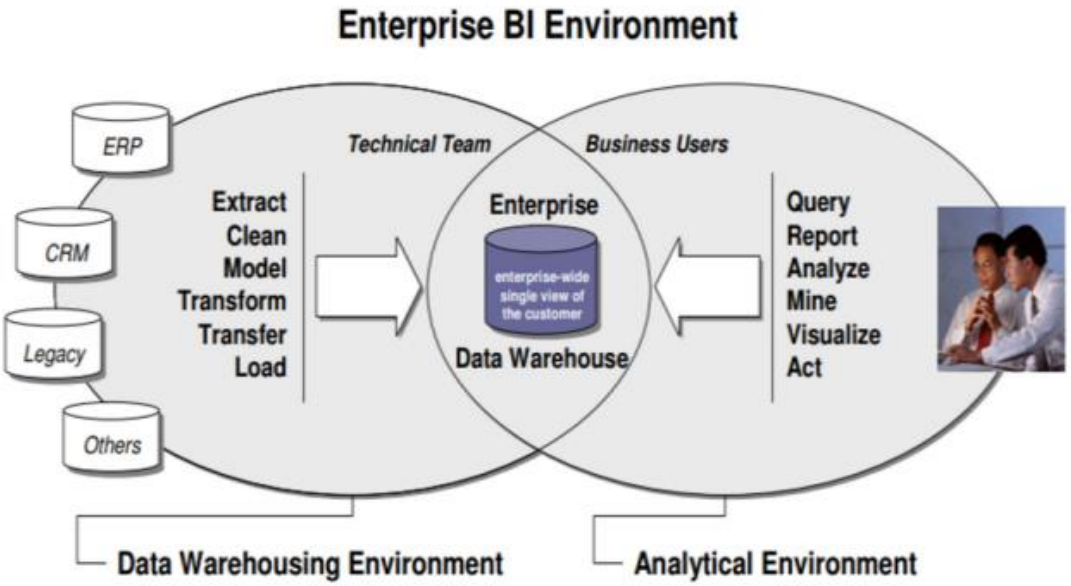


Figure 1. A typical enterprise BI environment.

These references demonstrate the growing significance of Big Data integration in BI, showcasing its diverse applications in economics, public sector decision making, and specifically, in the domains of closed funds management, education, healthcare, and public housing. Both the research shed light on the limitations of integrating data science with online analytical processing (OLAP) and reporting technology, specifically in terms of merging predictive modelling and decision-making functionalities.

2.1 Motives for this Real Estate Project

Data analytics is applied to this real estate project have several compelling reasons:

1. **Market Analysis:** Real estate data analytics offers perceptions into industry trends, assisting financiers and developers in making wise choices. We can use it to recognise new markets, evaluate supply and demand dynamics, and foresee price changes.

2. **Property Valuation:** By taking into account variables like location, property attributes, previous sales data, and market conditions, data analytics helps in valuing real estate. To appropriately price properties and maximise returns, this is essential.
3. **Risk assessment:** Investments in real estate developments are frequently significant. By examining market volatility, economic indicators, and other variables that may have an impact on a project's success, data analytics assists in risk assessment and mitigation.
4. **Property Development:** Data analytics helps real estate developers choose locations, make design choices, and identify the most profitable property kinds in particular regions. Forecasting building costs and schedules might also benefit from it.
5. **Sales and marketing:** Target demographics, desired property attributes, and efficient advertising channels are all identified through data analytics, which then guides marketing tactics. It supports client relationship management and lead generating.
6. **Cost management:** Examining operations and construction costs helps real estate projects maintain budgets, cut waste, and control costs.

Chapter 3: Literature Review

In Canada, the use of BI and ML in the real estate industry has been more popular recently. These technologies are being used more frequently by academics and professionals to improve risk assessment, investment strategies, market analysis, and other decision-making processes. Stakeholders may improve portfolio management, find investment possibilities, recognise market trends, and limit risks by combining these cutting-edge methodologies with real estate data.

This review aims to identify the key factors of Canadian real estate market shaping the market, highlight challenges and opportunities, and identify research gaps for future investigation.

3.1 Factors Influencing the Canadian Real Estate Market

3.1.1 Economic Indicators

Several studies have looked at how economic indicators like GDP growth, employment rates, and interest rates relate to their effects on the Canadian real estate market (Agarwal, 2022) . The demand and supply dynamics of the market are significantly shaped by these factors. Some of the key economic indicators that significantly effects the Canadian real estate market are:

1. **Interest Rates:**
   Mortgage rates are directly impacted by changes to the Bank of Canada's benchmark interest rate. By making borrowing more accessible, lower interest rates can increase demand for real estate, while higher rates can reduce it.
2. **Employment and Income Levels:**
   Potential homebuyers' purchasing power is strengthened by a healthy job market and rising salaries. In contrast, declining employment and sluggish income growth may reduce demand.
3. **GDP growth:**
   The rise of the gross domestic product (GDP) is a sign of the health of the economy as a whole. Higher GDP growth typically denotes more robust economic activity and, by enhancing consumer confidence, can have a favourable effect on the real estate market.
4. **Consumer Confidence:**
   A big factor is how consumers feel about the economy and their own money. Spending, notably on real estate, is encouraged by high consumer confidence.
5. **Population Growth and Demographics:**
   Immigration and natural population growth are the main causes of rising housing demand. Demand for various kinds of properties is also influenced by demographic changes like the ageing population or the influx of millennials.

3.1.2 Government rules and Policies

The market is substantially impacted by government rules and policies. According to studies (Canada, n.d.) and (estate, 2022), policies relating to mortgage rules, taxation, zoning, and housing affordability have an impact. Foreseeing market developments and their ramifications for different stakeholders, it is essential to understand the effects of these policies.

1. **Housing Affordability and Demand:**
   Government restrictions and lending criteria, for example, have a direct impact on housing affordability and demand. Potential homeowners' purchasing power may be constrained by strict mortgage stress tests that were implemented to ensure that borrowers could afford increased interest rates. Similar to how down payment restrictions affect demand levels, they also limit who can enter the market.
2. **Foreign Investment Regulations:**
   The real estate market can be strongly impacted by government actions intended to limit foreign investment. Demand may be impacted by taxes on foreign purchasers and restrictions on foreign ownership, particularly in significant urban areas with high levels of foreign investment.
3. **Rental Regulations:**
   The dynamics of the rental market are influenced by government rental housing laws, such as rent restrictions and tenant rights. Investors' rental yields may be impacted by rent control policies, while the availability of rental properties may be impacted by laws that are tenant friendly.
4. **Urban Development and Zoning Policies:**
   Local governments' zoning and land-use policies have an impact on the type and density of development in various places. These regulations may affect real estate prices, the availability of particular housing options, and the general appearance of cities.
5. **Economic Stimulus Measures:**
   Interest rates and fiscal stimulus programmes are two examples of government economic policies that can have a domino effect on the real estate market. While fiscal policies can have an impact on the general health of the economy and consumer confidence, lower interest rates can make borrowing more appealing and promote real estate investment.

3.1.3 Demographic Shifts

Population growth, migration trends, and changes in household composition are only a few examples of the demographic changes that have a significant impact on the real estate market (Canada, n.d.) and (estate, 2022) . In the Canadian context, researchers have looked at how changing demographics may affect regional differences, property kinds, and housing demand.

1. **Aging Population:**

The population of Canada is ageing, as is the case in many wealthy nations. Retirement homes, assisted living facilities, and downsizing are in higher demand as the baby boomer population approaches retirement age. The supply of single-family houses may also be impacted by this trend if older homeowners choose to downsize or relocate to retirement communities.

2. **Millennials' Entry into the Market:**

The real estate market is significantly fuelled by the millennial generation. As millennials enter their peak years for purchasing a home, their preferences and lifestyle decisions are changing the kinds of properties that are in demand. The demand for condominiums and homes in walkable urban locations is impacted by the fact that many millennials place a high value on urban living, amenities, and sustainable features.

3. **Immigration Patterns:**

The real estate market is directly impacted by Canada's immigration rules as well as population growth. The demand for these kinds of homes in particular areas is

influenced by the fact that new immigrants frequently look for rental properties or starter homes. Furthermore, immigrants may bring cultural preferences that affect housing options, such as multigenerational households.

4. **Urbanization and Suburbanization:**

More individuals are relocating to big cities seeking amenities and career possibilities as a result of demographic changes. This increases demand for condominiums and apartments near urban cores. However, as families look for bigger homes and more acreage outside of cities, suburbanization continues.

5. **Household Formation:**

The types of homes in demand are impacted by changes in household composition. For instance, the demand for smaller flats and condos may increase as the number of one-person households rises, while the demand for larger homes in suburban areas may rise as families expand.

3.1.4 Market Dynamics

The Canadian real estate market is heavily influenced by market dynamics, including supply and demand dynamics, investor mood, and market liquidity (Canada, n.d.) and (estate, 2022) . Predicting market trends and spotting investment opportunities require an understanding of these forces and how they interact.

1. **Economic Conditions:**
   The real estate market is directly impacted by economic factors including GDP growth, employment rates, and inflation. A healthy economy may result in greater demand for homes and higher prices.
2. **Interest Rates:**
   Mortgage rates are influenced by changes in interest rates set by the central bank. Higher rates can limit demand while lower rates can increase it as borrowing becomes more accessible.
3. **Supply and Demand Dynamics:**
   Property prices are significantly influenced by how supply and demand for housing are balanced. Cities with a strong demand and a constrained supply frequently see quick price increases.
4. **Demographics:**
   Housing demand is impacted by demographic changes, such as population growth, ageing populations, and changes in household formation. Immigration and urbanisation are also important factors.
5. **Housing Market Speculation:**
   Investors buying real estate for short-term gains might result in speculation-driven price hikes, but speculation can also result in unstable markets.

3.2 Previous Studies on Real Estate using Data Analytics.

Recent years have witnessed substantial expansion and interest in the field of real estate data analytics, which has sparked a number of studies and investigations. The real estate sector is covered in a wide variety of these studies' subjects and applications. Here are some noteworthy subfields of study and case studies in the subject of real estate data analytics:

1. **Price Prediction and Validation:**
   The study is titled "Predicting House Prices with Machine Learning Models" and was written by Yu Bao, Yu He, Yu Wu, and Yu Fan.

Focus: A crucial component of real estate assessment, the application of machine learning models to anticipate house prices is explored in this study.

2. **Market Analysis and Trend:**
   Research: R. Moraes' "Real Estate Market Analysis Using Big Data Analytics"
   Focus: The study covers the use of big data analytics to examine and comprehend market patterns in the real estate industry, assisting investors and developers in making wise decisions.
3. **Risk Assessment:**
   The study is titled "Real Estate Risk Assessment Using Data Analytics" and was co-authored by He, S., Wong, R. K. W., and Xu.
   Focus: This study examines how data analytics can be used to evaluate market and financial risks as well as other hazards related to real estate transactions.
4. **Property Portfolio Optimization:**
   Study by Srinivasan, R., "Optimising Real Estate Portfolio Performance Using Data Analytics"
   Focus: By identifying underperforming assets and optimising allocations, the study investigates how data analytics might be utilised to improve the performance of real estate portfolios.
5. **Property Management:**
   "Data Analytics in Real Estate Property Management" by A. Ghahramani and P. Gohari.
   Focus: This study investigates how data analytics might enhance the processes involved in real estate management, such as predicting maintenance needs, maximising rental rates, and ensuring tenant happiness.
6. **Market Forecasting:**
   Research: "Real Estate Market Forecasting Using Machine Learning" by Chen, J., Zhang, & Chen X.
   The research's main focus is on using machine learning to forecast real estate market patterns and project future property values.

3.3 Data Integration and Warehousing

Various data sources, including property listings, transaction histories, market trends, and demographic information, should be integrated into centralised data warehouses, according to studies. These repositories make it possible to instantly obtain thorough data to aid in decision-making. (Chen, 2019)

3.3.1 Role of Data Warehousing

A recognised basic technology for storing and managing integrated real estate data is data warehousing. The necessity of well-designed data warehouse designs for enabling effective querying, reporting, and analysis is stressed by studies. In the literature, several data warehousing strategies—including traditional data warehousing and cloud-based solutions—are discussed along with their effects on the real estate industry.

3.3.2 Decision-Making Support

The use of integrated data to enhance decision-making at various stages of real estate projects is highlighted in the literature. Real-time access to vital information made possible by data warehousing assists stakeholders in determining risk tolerance,

pricing plans, and investment strategies. Additionally, integrated data makes scenario analysis easier and aids in resource allocation optimisation.

3.3.3 Process Efficiency and Automation

The literature emphasises how data warehousing and integration may speed up business operations in the real estate industry. Automation can enhance efficiency, save operational costs, and minimise errors in property management, lease administration, and financial reporting. (Li, 2020)

3.3.4 Challenges in Implementation

While the potential advantages of data warehousing and integration are acknowledged, implementation difficulties are also covered. These include organisational problems like change management and resistance to technology adoption as well as technical challenges like choosing the right technologies and guaranteeing data security. (Ma, 2018)

3.4 Business Intelligence

Researchers emphasise how BI technologies can turn raw data into useful insights. Real estate professionals may monitor market trends, manage key performance indicators, and spot investment possibilities thanks to dashboards, data visualisation, and reporting tools. (Tanwar, 2017)

3.4.1 Customer-Centric Approaches

Researchers emphasise how BI may improve consumer experiences. Real estate companies can better target their services, provide individualised property recommendations, and increase overall client involvement by analysing customer data.

3.4.2 Operational Efficiency

By automating repetitive procedures, easing property management, and optimising resource allocation, BI solutions optimise operational processes. This effectiveness results in lower costs and more production.

3.4.3 Decision Support Systems

A common subject is the creation of decision support systems that incorporate BI and real estate data. These tools let real estate experts evaluate investment opportunities, forecast market trends, and control risks.

3.4.4 Implementation Strategies

Researchers stress the significance of a carefully thought-out implementation approach for BI in the real estate industry. This entails connecting technology to organisational objectives, promoting a data-driven culture, and making sure there is sufficient user training. (Smith J. A., 2020)

3.5 Machine Learning

Due to its potential to improve decision-making, market analysis, and predictive modelling, the incorporation of machine learning (ML) techniques in the real estate sector has drawn considerable attention. Researchers have looked into a variety of aspects of ML in real estate, from demand forecast to risk assessment and property appraisal (Kakkar, 2019) .Some of the research papers which were published on Price predictions on Real estate dataset are (Bao, 2018) , (Şeker, 2018) and (Aziz, 2019).
(Bao, 2018) examines the use of different machine learning models, such as support vector regression, random forests, and linear regression, to forecast home prices. It goes over approaches for feature selection, data preprocessing, and model evaluation. (Şeker, 2018) uses multiple regression analysis to forecast real estate values. It talks about creating regression models for price prediction and choosing pertinent predictors. Whereas, in (Aziz, 2019), the accuracy of four distinct home price prediction models—linear regression, random forests, gradient boosting, and neural networks—is compared. It offers information about how well these models work in the context of real estate valuation.

3.5.1 Price Prediction and Valuation

Numerous studies concentrate on applying machine learning algorithms to forecast real estate prices. These models use historical sales information, property traits, and neighbourhood features to precisely estimate property values. To improve forecast accuracy, researchers have used ensemble approaches, neural networks, decision trees, and regression algorithms.

3.5.2 Demand Prediction

By examining variables including population growth, economic indicators, and rates of urbanisation, ML models have been used to estimate housing demand. Developers and investors can use these forecasts to find prospective development locations.

3.5.3 Risk Assessment

Researchers have looked into how ML can evaluate the risks connected to real estate investments. In order to estimate the possibility of default, algorithms can analyse previous data, which helps lenders make wise lending decisions.

3.5.4 Market Trend Analysis

To find patterns and trends in the real estate market, machine learning algorithms analyse vast amounts of data. This helps stakeholders comprehend market dynamics, spot new trends, and adjust strategy as necessary.

3.5.5 Property Management and Maintenance

To predict maintenance requirements for real estate facilities, ML-based predictive maintenance models have been created. This cuts downtime and operational expenses. (Kakkar, 2019)

Chapter 5: Methodology

This study offers a thorough methodology for combining business intelligence (BI) and machine learning (ML) approaches to improve data handling and decision-making in the real estate sector. The suggested technique intends to harness the power of BI tools and ML algorithms to analyse massive real estate data volumes, derive valuable insights, and enable informed decision-making by diverse stakeholders, including investors, developers, and property managers. Data collection and preprocessing, feature engineering, model selection and training, and the creation of decision support systems are all integral parts of the methodology. It also deals with issues like data quality, privacy, and scalability. The effectiveness of the proposed methodology is demonstrated through a case study involving real-world real estate data, showcasing its potential to improve decision-making processes and optimize resource allocation in the dynamic real estate market. Overall, this methodology provides a valuable framework for leveraging BI and ML in the real estate domain, enabling stakeholders to make data-driven decisions and gain a competitive edge in the industry.

5.1 Research Design

The underlying technologies that serve as the technical foundation for the proposed BI system were discussed in (Sherman, 2014) and their relationships are depicted in Figure 2 for the proposed BI system. The following crucial considerations should be kept in mind because technology and tool selection directly affect the quality of the data:

- ETL operations extract and load the data and load it into destination.
- Web crawling tools within Python selenium library retrieve data from real estate website. (Saini, 2016)
- Web scrapping will be done with the help of hidden API through python programming.
- Python will be used to build up PostgreSQL connectivity in order to save all the website data that was scraped.
- To make the data comprehensible for BI tools, Pentaho's data integration tool will be used. (Campoli)
- Data that has been cleaned and processed will then be loaded into various PostgreSQL tables.
- The implementation of data analytics capabilities uses Python-based machine learning methods.
- Finally, PowerBI will provides the dashboards and visualization instruments for end users.
- On the other hand, Machine Learning will be performed as well after cleaning of data in PostgreSQL.
- Machine Learning algorithms like Linear Regression, XGBoost regression (SALAM FRAIHAT, 2021) and much more will be applied on the Data which is provided by PostgreSQL.

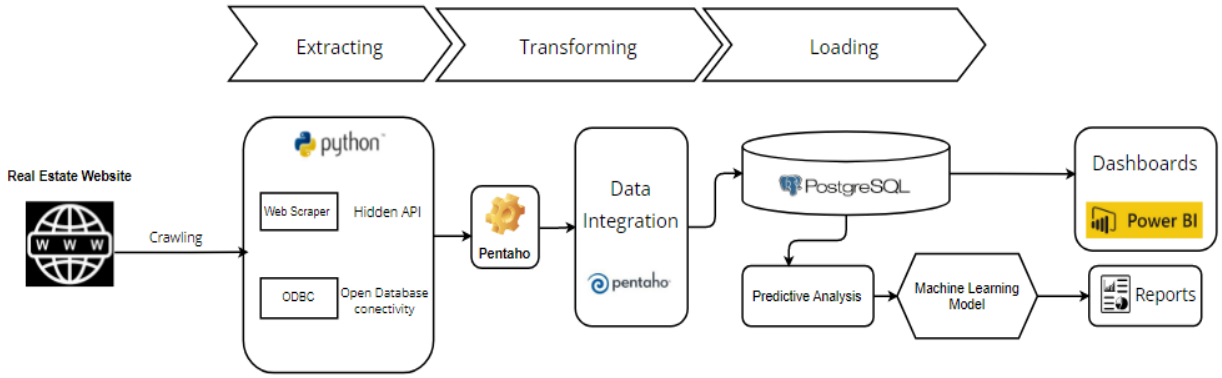


Figure 2. An enterprise BI environment.

5.2 Data Requirements

Data quality is affected by a variety of things, such as different data sources, tool constraints, and human mistake. Several actions that improve data quality measures, such as completeness, conformance, accuracy, timeliness, consistency, and integrity, can increase data reliability.

The important components of the BI system are data collection, preparation, and loading, which determine the system's capacity to serve its intended function. The quality of the final system is directly impacted by the data source, formats, cleansing, normalisation, and other factors of the ETL process (Sherman, 2014)

These procedures are thoroughly explained in this section.

5.2.1 Phase 1: Data Collection

The suggested approach calls for a data feed from numerous sources and formats, which may include both unstructured data from web pages and other documents and structured data from legacy systems and databases, including Excel and CSV. Specifically designed web crawlers are used to gather web data by scanning source websites and extracting the necessary data. We can utilise the Selenium library, for instance, in Python development environments. Libraries like the Natural Language Toolkit (NLTK) for unstructured documents are also available in Python development environments.

In this project, Extraction of data will be performed from the Canadian Real Estate website Realtor.ca using web crawling capabilities in the Python Selenium package. Scraped data will be saved as a Dataframe, which will be then used to transport the data to a PostgreSQL database using an ODBC connection as shown in figure 2.

5.2.2 Phase 2: Data Cleaning, Transformation, and Integration

This process requires that you:

- Ensure that data formats are as expected/desired, which may involve converting text fields to numerical values (prices).
- Fix missing and noisy values with replacement or removal.
- Ensure that data formats are as expected/desired.
- Support readiness for data integration with well-known tools like Pentaho and SSIS.

The project's proposal states that the scraped data that is currently kept in the Raw table would be converted into clean data and kept in the Master table. The associated data from that master table will be saved in additional tables that will be made later. Figure 2 shows the architecture.

5.2.3 Phase 3: Data Analytics

Pentaho and PostgreSQL (Campoli) will serve as the data stream for the Business Intelligence dashboard tool like Power BI or Tableau to satisfy this need.Also, Machine learning will be applied accordingly to predict the house price with the help of different algorithms like Linear regression, XG Boost regression (as previously worked by (SALAM FRAIHAT, 2021)) and much more. Figure 2 shows the detailed architecture.

5.3 Predictive Analysis

Property prices are a key macroeconomic indicator for national economies, and changes in property values have an impact on market boundaries, investment opportunities, and investment decisions. Location, age, size, and proximity to services and facilities are just a few of the variables that affect a property's value. In the past, there was no formal certification procedure, no established set of acceptable criteria, and the traditional methods of estimating property prices were based on sale price, land cost, and building costs. This is further complicated by the fact that many real estate brokers and investors tend to be ignorant of the relevant factors involved in the thorough evaluation of real estate prices. By considering all factors in a way that might increase market efficiency, predictive analytics can make a good contribution to real estate price estimation and offer far more thorough assessments to all stakeholders. With a good set of correlated indicator data, which can be variables that represent land or property features like size, age, and location, machine learning algorithms, such as random forests, neural networks, and linear and logistic regression, are capable of predicting real estate prices, as shown in Figure 3.

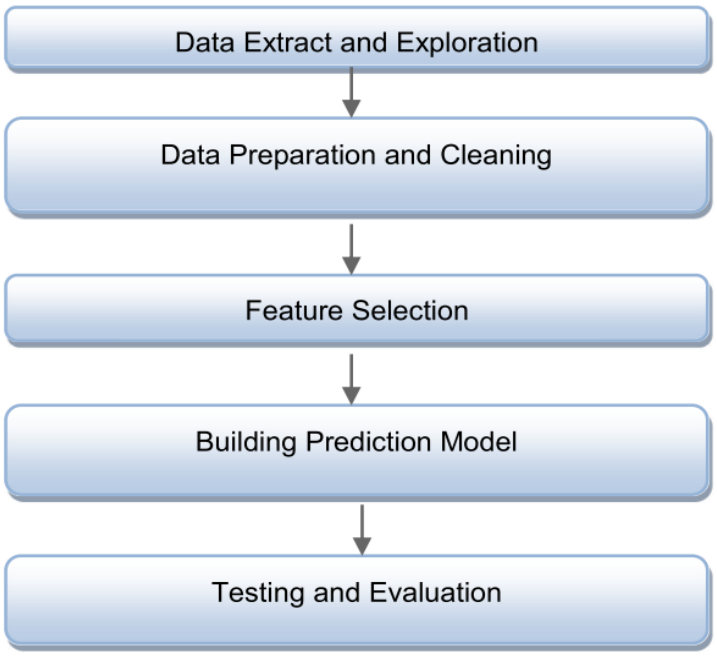


Figure 3. An enterprise BI environment.

5.4 Project timeline

This research proposal has shown that the main difficulties in this area of study are due to the traditional process mining's inherent errors and lack of precision. Process mining is also being looked into as a potential solution to these issues, bridging the gap between data handling and data mining approaches and providing more meaningful data for decision-making. The research project's schedule is shown in the timeline that follows in Figure 3 in order to accomplish the stated aim and objectives.

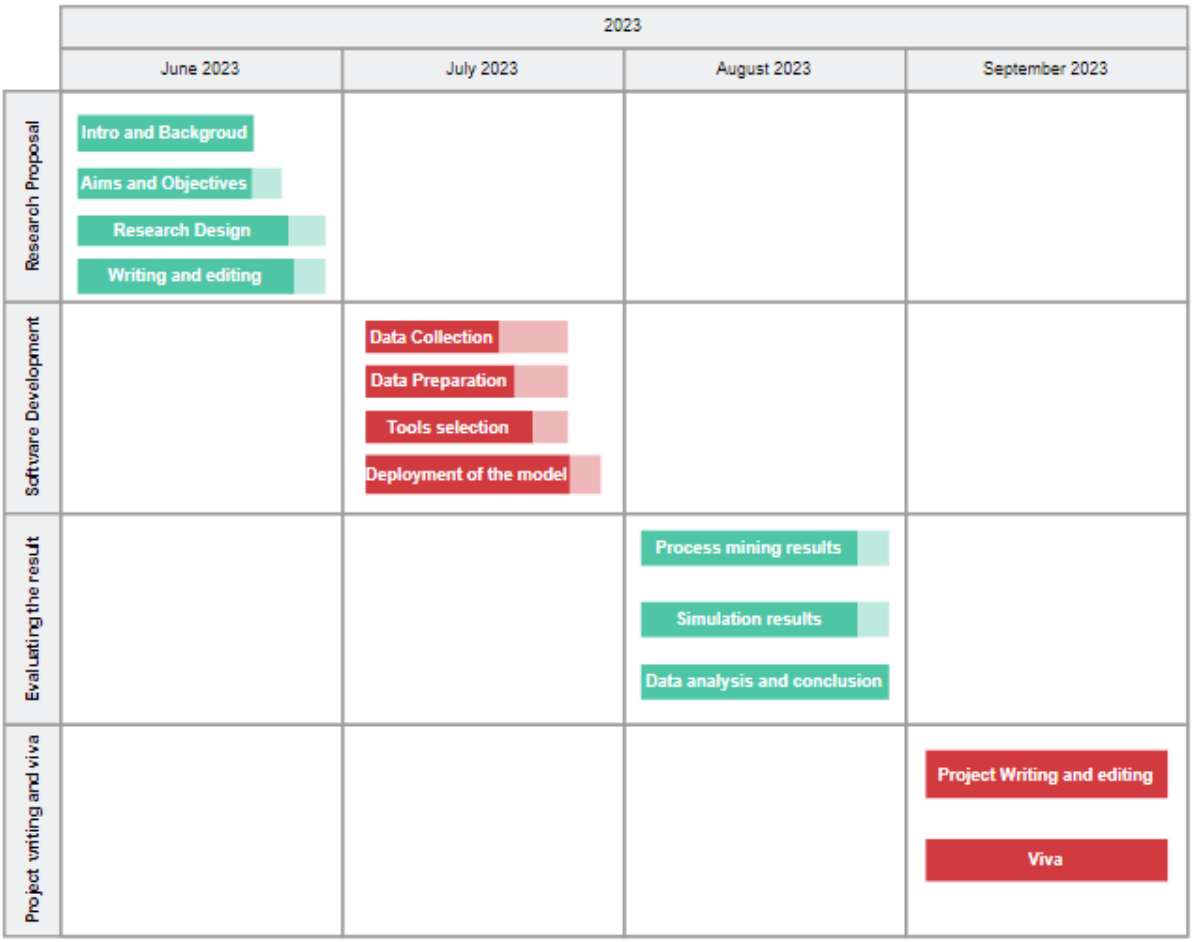


Figure 4. Project Timeline

Chapter 4. Design and Development of the System

This chapter explains the design, development, and deployment of the model that extracts, transforms, and loads data into separate tables that are subsequently used to generate reports. Additionally, a dataset that is being cleaned is taken out of the database, and a predictive model is run for price prediction analysis.

4.1 Data Extract, Transform and Load (ETL)

Every data-driven real estate project needs to include the Extract, Transform, and Load (ETL) process. Data must be extracted from a source, formatted appropriately, and loaded into a data warehouse or repository for analysis. The ETL procedure for this real estate project is described in detail below.

4.1.1 Data Extraction

Data Extraction is one of the critical steps in ETL (Extract, Transform and Load). Extraction of data from a source includes Accessing the data source which could be databases, files, Web APIs or even Web scrapping. In this project, data from a Canadian website called Realtor.ca has been accessed as the data extraction on website is legal for performing analytics. If we are unable to access real estate data from a website, this project won't start at all unless I begin working on dataset provided online from Kaggle.com.

As the realtor.ca website has provided access to its data retrieval API, which is accessed using python, various steps have been followed in order to do web scraping with the aid of a hidden API using python programming on a Jupyter notebook.

- Requesting data from website by web scrapping.
- Checking status of the data.
- Data which was need is append into a Panda’s Dataframe.
- Data loading into PostgreSQL with the help of ‘SQL AlChemy’

A query pops into your head. Why did you choose to collect data from a Canadian real estate website rather of a UK real estate one? I was unable to access the information that was provided on the real estate websites for the United Kingdom, in accordance with their rules and regulation. Only if someone pays for it and acquire ownership of the data, the website's owners are not permitted to make the data publicly available to anybody.

4.1.2 Data Transformation

Data transformation is the process of taking the uncleansed raw data that has been retrieved from source, transforming it into a format appropriate for analysis, reporting, or putting into a data warehouse or a Database. In this project, Data which was extracted from realtor.ca as mentioned in 4.1.1 was loaded into PostgreSQL as in Raw format. The Table in which the raw data is stored is name as ‘real_estate_scrap_data’ in PostgreSQL database.

4.1.2.1 ERD Diagram

An ERD diagram has been planned for data modelling that defines the structure and relationships within the PostgreSQL database as shown in Figure 5. ERD diagram shows us the Database design, Visual Clarity and normalization of data inside a PostgreSQL database.

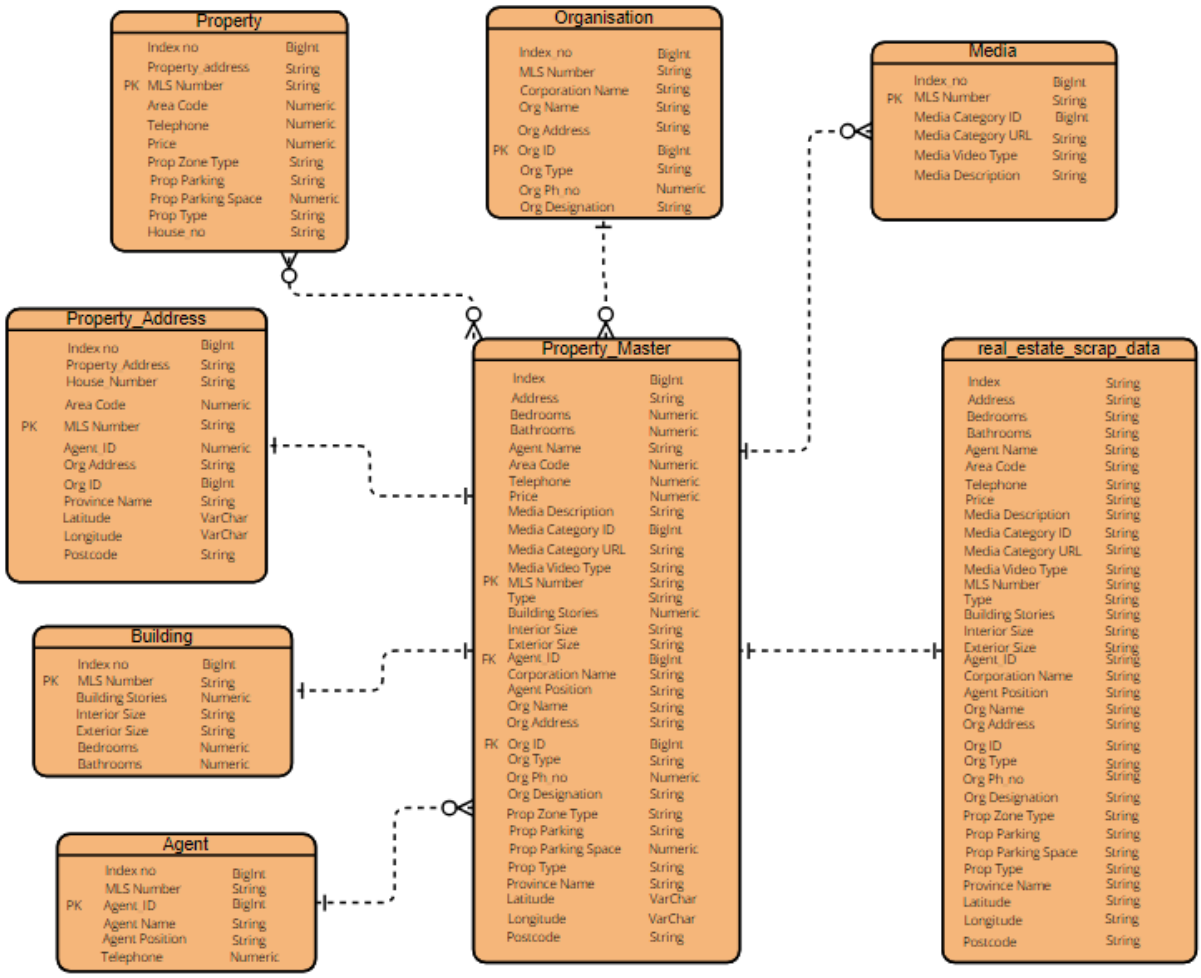


Figure 5.ERD Diagram of PostgreSQL Database.

Figure 5 illustrates the degree to which each entity is tied to the others. The aforementioned Raw data is kept in 'real_estate_scrap_data'. This raw data table and Property_master are directly related to one another forming on-to-one relationship. Data that has been cleaned and normalised from the Raw table, which was scraped using Python, is contained in the Property master table. Due to the fact that this table is the master and provides data to each unique entity separately, Property_master has a different relationship with the other entities. Agent Entity has a one-to-many relationship since each agent has numerous attributes that are stored in the Entity Property Master. Detailed information on the building is provided by the building entity for the building that has been listed on the website. Building and Property_master have a one-to-one relationship that is identical to the relationship between Property_Master and the raw table ‘real_estate_scrap_data’. The property_address table, which gives the detailed address of a property in the property_master table, has a similar one-to-one relationship. The Property table and Property_master table in the ERD diagram above have a special many-to-many relationship. Since both offer the same data, the

property table makes things easier by compiling only the pertinent information as opposed to property_master, which has all the accurate information. To keep track of all the details of the organisation that an agent or piece of property belongs to, an organisation table is built. Organisation and property_master have a one-to-many relationship since one organisation has several properties listed on the website. Last but not least, the Media entity contains the website's listed property's electronic media information. Due to the large number of photographs and videos that are provided for a single property on the website along with the media details like description and type of media, the Media table has a many to one relationship with the Property_master table.

4.1.2.2 Pentaho Data Integration

The open-source ETL (Extract, Transform, Load) programme Pentaho Data Integration, sometimes referred to as Kettle, is useful for managing and transforming real estate data. It has a variety of features and skills that enable it to manage the intricate requirements of real estate data integration and analysis.

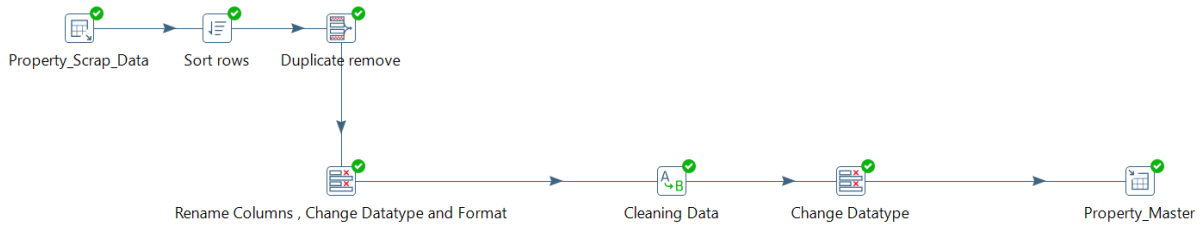


Figure 6. Pentaho transformation of Raw data to Clean and normalized Data

Figure 6's transformation demonstrates that there are numerous irregularities in the raw data that was extracted and placed in the ‘real_estate_scrap_data’ table. The principal column "mls number," which is used by working real estate brokers to communicate information about homes up for sale, has duplicate entries in the raw data that have been stored. A few columns need to have their names changed in order to better reflect the facts. The data had to then be cleaned up because certain columns contained some letters or alphabet that wasn't supposed to be there and was having an impact on the column's data. All of the columns in a raw table had string datatypes, which needed to be changed depending on the data that each column contained. In response, the datatypes were modified. Finally, the cleaned data is then stored in Property_master table.

4.1.2 Data Loading

The last phase in the ETL (Extract, Transform, Load) process is data loading, which involves putting the cleaned and converted data into a target location such a data warehouse, database, or other storage systems. The availability of the modified data for analysis, reporting, and other business operations depends on this stage.

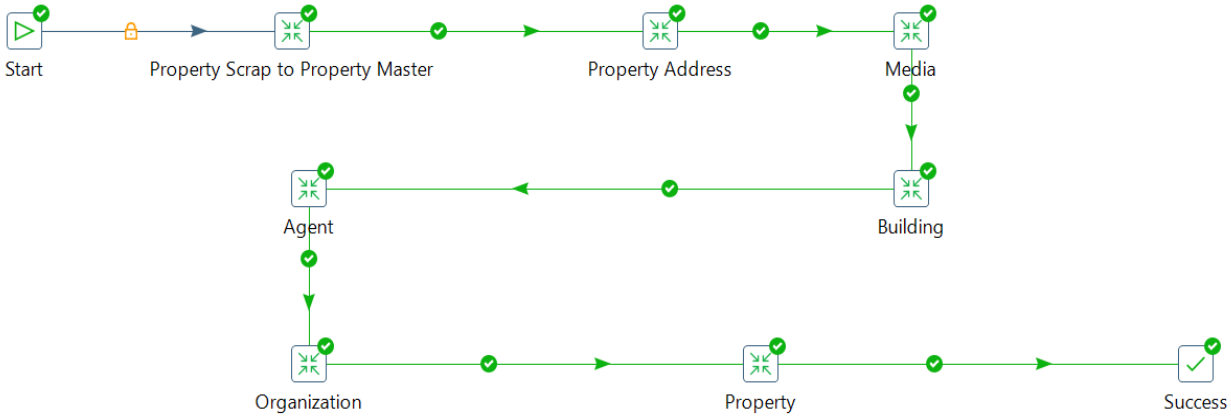


Figure 7. Clean data transferred to its relevant tables.

Figure 7 shows how to move cleaned-up data that was previously stored in Property_master into the appropriate tables. The essential real estate data is stored in a variety of tables, including information on buildings, such as the 'building' table in the PostgreSQL database stores the number of stories, internal size, and external size. Similar to this, information about agents is kept in the agent table, information about organisations is kept in the organisation table, information about properties is kept in the Property_address table, and information about media is kept in the Media table. The property table simplifies things by gathering only the relevant information rather than the property_master, which contains all the accurate information.

4.2 Business Intelligence

In the context of a real estate dataset, business intelligence (BI) includes the use of data analysis and visualisation tools to derive actionable insights and enable data-driven decision-making within the real estate sector. Data is entered into the PowerBI Desktop application once it has been loaded into all the necessary tables where it belongs and is then visualised to provide helpful insights for the decision-making process. These visualisations are used for a variety of purposes, including market analysis, forecasting, financial analysis, reporting, and performance monitoring. The reports that were generated from the data that was extracted, processed, and loaded into the appropriate tables are displayed in the chapter that follows.

4.3 Machine Learning

Machine learning is an organised approach to developing prediction models and utilising data to guide decisions. In the literature review above, machine learning is briefly defined. The basic steps of this project that make up a typical machine learning workflow are listed below:

4.3.1 Data Collection

Gather relevant data for our problem from various sources. This may include databases, APIs, web scraping, sensor data, and more. Ensure that the data is clean, complete, and representative of the problem.

The data used in this project was gathered from a Canadian website named realtor.ca. The website's owner made the data openly accessible to all users for data analytics.

With the use of Python programmed in a Jupyter notebook, data from the realtor.ca website is scraped using the Hidden API. The real estate data is carefully chosen and scraped from realtor.ca before being saved in a dataframe. It is then placed in a table called "real_estate_scrap_data" in a PostgreSQL database as raw data. This process is shown in above figure 2.

4.3.2 Data Preprocessing

Data preparation is an essential stage in the workflow for data analysis and machine learning. In order to prepare raw data for analysis or for training machine learning models, it must be cleaned, transformed, and organised.

1. Data Cleaning

Instead of cleaning data with Python, it is cleaned in the PostgreSQL database. The raw data which is stored in PostgreSQL Database is first cleaned by removing data duplication from the primary key which is mls_number (Multiple listing services) which are created and maintained by cooperating real estate professionals to share information about properties on the market. The columns will need to be renamed as the users may not comprehend them by default. As a result, they are given new names based on the information in the columns. Additionally, formats and datatypes are altered based on the information contained in tables, while string datatypes are used by default for all data storage. Some of the columns' datatypes won't update since they contain anomalies such If the datatype is numeric, there should only be numeric data present; nonetheless, some garbage values or an alphabet that makes no sense are also there. Therefore, the data is cleaned by eliminating those useless values or Alphabets, and then the datatype is modified. Finally, the now-clean data is put into a main table called "Property_Master". The process is shown in Figure 6 above. In order to employ machine learning techniques, the data from the property_master table is stored in a CSV file and then sent in to juypter Notebook.

2. Exploratory Data Analysis (EDA)

In the machine learning pipeline when using real estate data, exploratory data analysis (EDA) is a critical stage. To get insights, spot patterns, and make wise choices about data preprocessing and model construction, it entails analysing and comprehending the dataset.

- Data which is cleaned in PostgreSQL is then analyzed by checking the shape and size of the dataset.
- Data may contain missing or null values in columns which is manipulated by the command called dropna() from Panda’s Library.
- Learning the dtypes of columns' and how many non-null values are there in those columns.
- Getting the statistical summary of dataset
- Visualizing the correlations between numerical variables in the dataset which is as shown below figure 8.

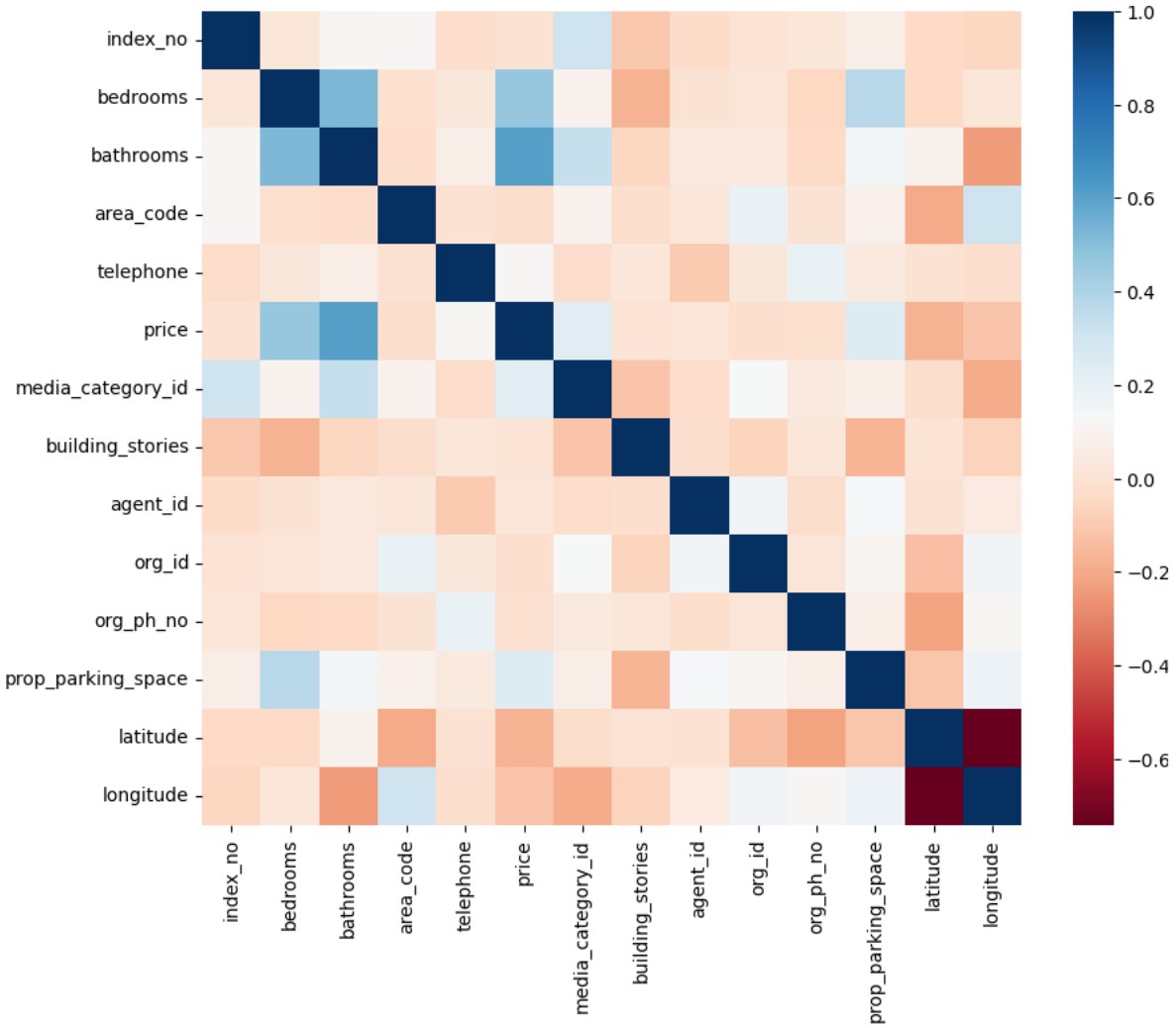


Figure 8. Correlations between Numerical variables

3. Feature Selection

Choose the input variables for my model that are the most pertinent features. It is possible to improve the performance and comprehension of the model by adding new characteristics. We are selecting numerical features which have more than 0.40 or less than -0.40 correlation rate based on Pearson Correlation Method—which is the default value of parameter "method" in corr() function. As for selecting categorical features, I selected the categorical values which I believe have significant effect on the target variable such as Prop_parking_space. The crucial columns that are connected to one another are Price, Bedrooms, and Bathrooms. The results of machine learning models can then be affected by checking the missing values in these columns which is why I remove any missing values in these columns. Figure 9 shows the visualization of these 3 tables.

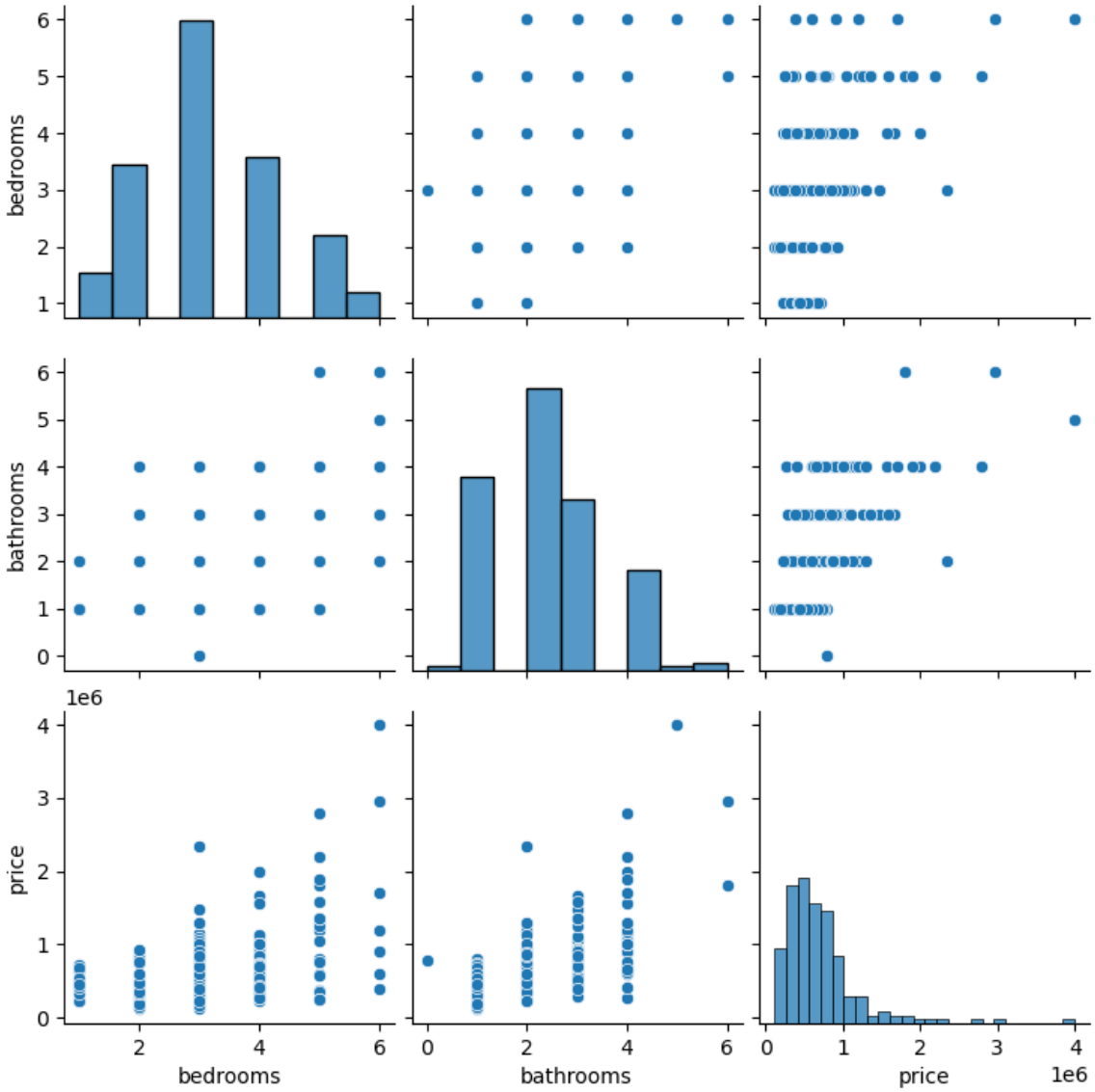


Figure 9.Data visualization of correlated numeric columns.

4. Data Standardising and Splitting

The processes of data standardisation, which entails scaling numerical features to have a mean of 0 and a standard deviation of 1, and data splitting, which entails dividing my dataset into training and testing subsets to assess your model's performance, are crucial steps in getting a real estate dataset ready for machine learning.

- **Split and Encoding**

  The dataset has been divided into X and Y Dataset, where X is the bedrooms and bathrooms data without Price and Y is the Price data solely.
  One-Hot Encoding has been applied in which Encoding the categorical features in X dataset.

- **Standardising the data**

  Standardizing the numerical columns in X dataset. StandardScaler() adjusts the mean of the features as 0 and standard deviation of features as 1. Formula that StandardScaler() uses is as shown in figure 10.

$$z = \frac{x - \mu}{\sigma}$$

μ = Mean

σ = Standard Deviation

Figure 10. Formula of StandardScaler()

- **Train-Test Split**

The train-test split is a crucial phase in the model generation process in machine learning. The dataset will be split into two distinct subsets, one for training the model and the other for assessing its performance. When determining how effectively the models will generalise to new data, this division is crucial. Splitting dataset into 2 chunks where 20% of the data is stored in test and 80% of the data is stored in for training the models.

4.3.3 Data Modelling

For a real estate dataset, data modelling in machine learning entails building a predictive model that can make intelligent choices or produce insights based on the data. To assess and enhance these models, it is crucial to consider a number of fundamental ideas and metrics:

1. **Mean Absolute Error (MAE):** The mean absolute difference between the expected and actual values is measured by MAE. A lower MAE denotes higher performance and offers a clear insight of prediction accuracy. It does not, however, penalise significant errors as severely as other metrics.
2. **Mean Squared Error (MSE):** MSE computes the mean of the squared discrepancies between the expected and actual values. Larger errors have a greater impact when the errors are squared, which increases the sensitivity of MSE to outliers. A lower MSE value denotes improved model performance.
3. **Root Mean Squared Error (RMSE)** is the square root of Mean Square Error (MSE). Compared to MSE, it is easier to read because it is in the same unit as the target variable. Similar to MSE, lower RMSE values signify better accuracy, and it is similarly outlier sensitive.
4. **R-squared (R2) Score:** R2 gauges how much of the variance in the target dependent variable (the target) can be attributed to the model's independent variables (the features). A better fit is indicated by higher values, which range from 0 to 1. A high R2 indicates that the model adequately accounts for the data's variation.
5. **Cross-Validation:** By dividing the dataset into various subsets (folds), cross-validation is a technique used to evaluate a model's performance. Each fold serves as both a training and a validation set since the model is trained and assessed several times. This improves the model's performance estimate and aids in the detection of overfitting.

Linear regression, Ridge regression, Lasso regression, Elastic Net, Support vector machines, Random Forest Regressor, XG Boost Regressor, and Polynomial Regression are the primary algorithms used in data modelling for a real estate dataset. The real estate data produced by these algorithms shows excellent outcomes. Following that, I compared them and plotted them appropriately, as is demonstrated in the following chapter showing the results.

Chapter 5. Discussion of Results

The project's findings are quite exciting because many new insights were discovered. reports from business intelligence to predictions from machine learning. Here are the outcomes I've been working towards for the past three to four months after all the effort described in the design and development.

5.1 Business Intelligence

Numerous reports are produced, each of which reveals a distinctive insight.

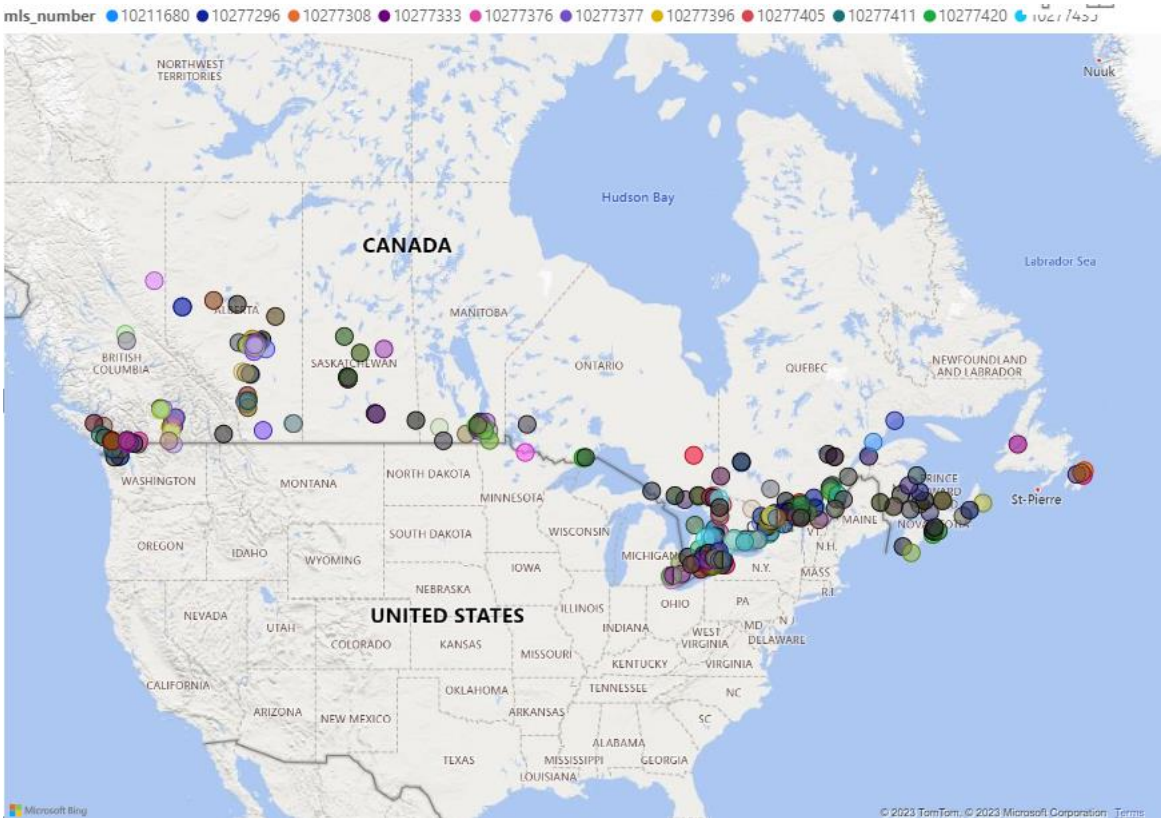


Figure 11. Property Geolocation on Map

The report is shown in Figure 11 which provides us with a visual representation of location by displaying all of the property's geolocation. As PowerBI has a useful function, clicking on one of the geolocations will display the details next to it.

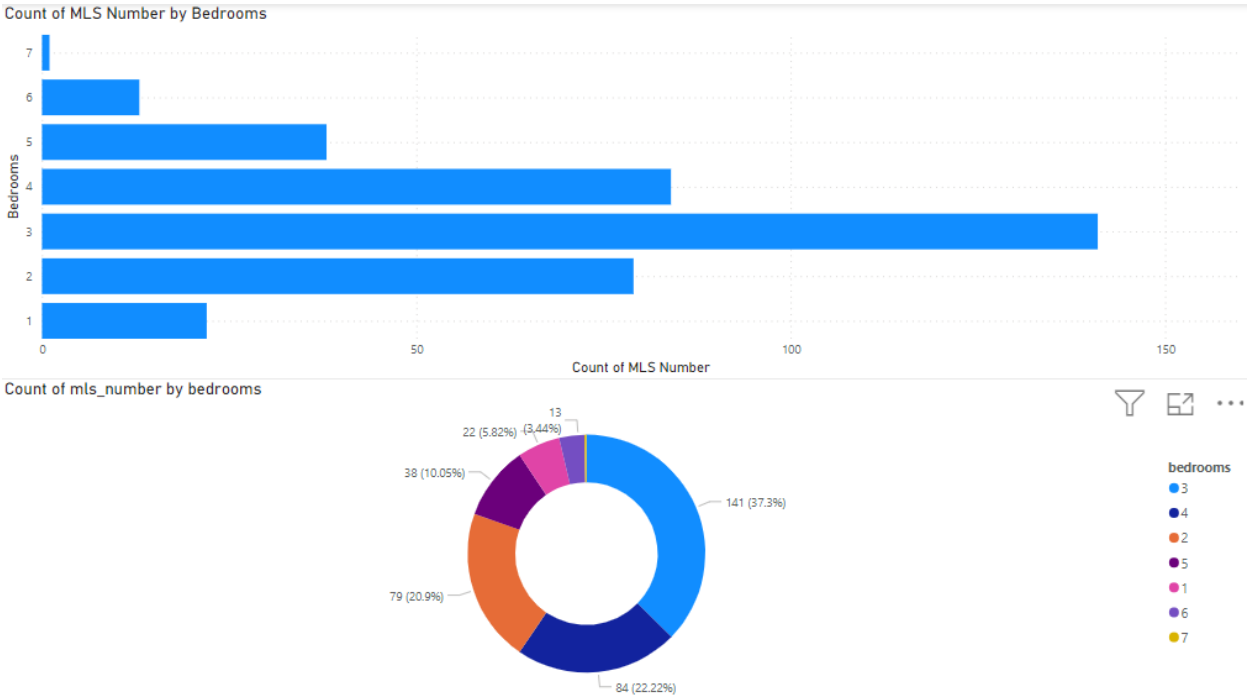


Figure 12. Data visualization of Bedrooms in number of properties

Figure 12 shows properties that are having bedrooms as 3 Bedroom Property is more in quantity than the others. At 141 out of the total Property listed, 3 bedrooms had the highest Count of MLS Number and was 14,000.00% higher than 7 Bedrooms Property, which had the lowest Count of MLS Number at 1. 3 Bedrooms Property accounted for 37.30% of Count of MLS Number. Across all 7 Bedrooms, Count of MLS Number ranged from 1 to 141 of Property listed on Realtor.ca.

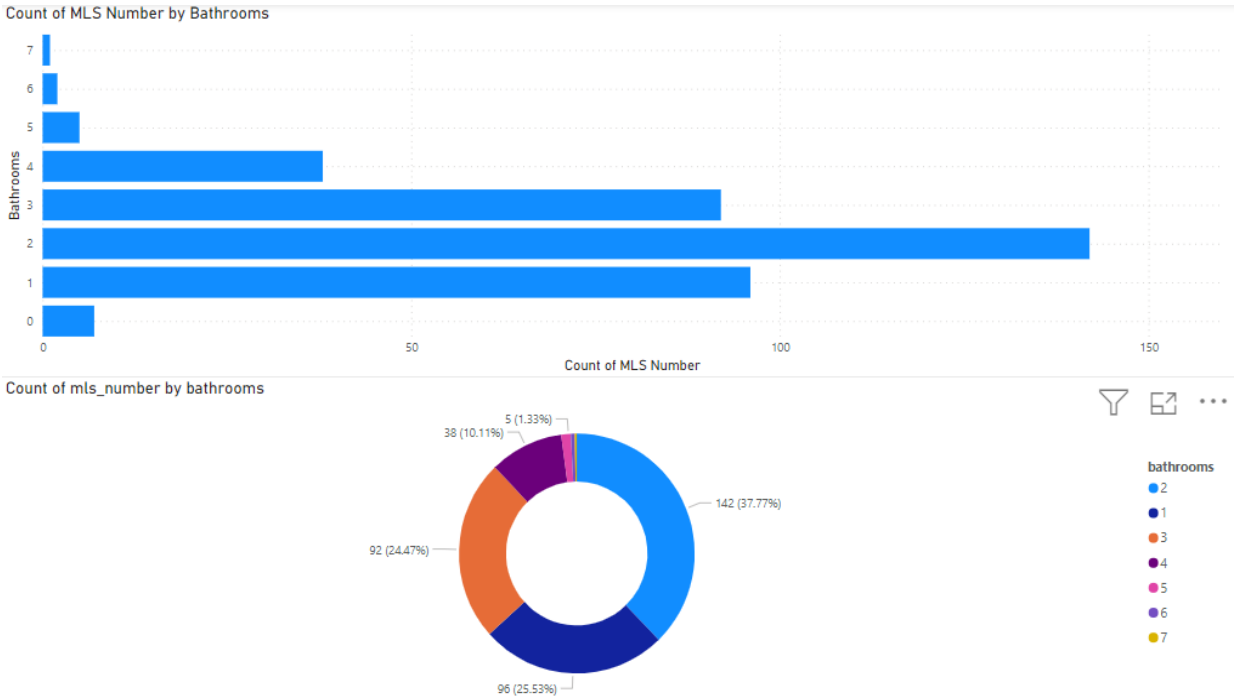


Figure 13. Data visualization of Bathrooms in number of properties

Figure 13 shows properties that are having Bathrooms as 2 Bathroom Property is more in quantity than the others listed on website. At 142 out of the total Property listed, 2 Bathrooms Property had the highest Count of MLS Number and was 14,100.00% higher than 7 bathrooms Property, which had the lowest Count of MLS Number at 1 property listed that contains 7 Bathrooms. 2 Bathrooms Property accounted for 35.15% of Count of MLS Number. Across all 9 Bathrooms, Count of MLS Number ranged from 1 to 142.

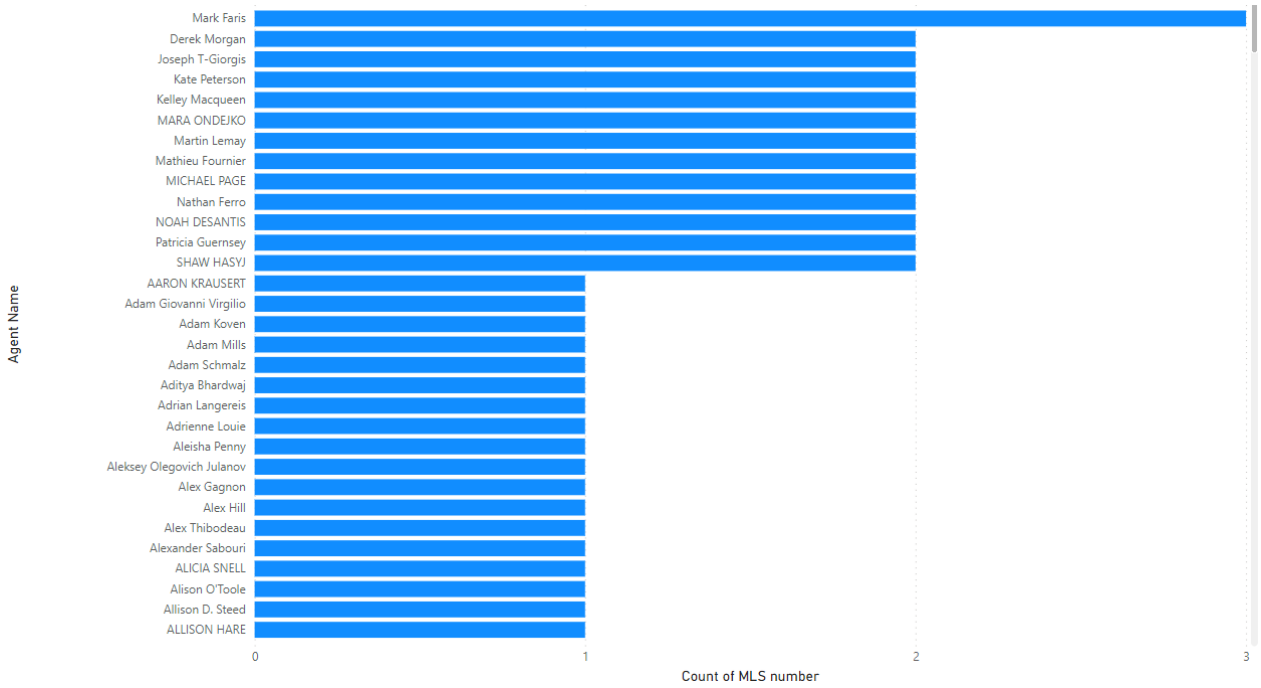


Figure 14. Count of MLS number by Agent name

Figure 14 shows those agent names who have listed Properties on Realtor.ca website. Agent Mark Faris had the highest Count of MLS number at 3, followed by Derek Morgan and Joseph T-Giorgis, which tied for second at 2. Mark Faris accounted for 0.74% of Count of MLS number. Across all 390 Agent Name, Count of MLS number ranged from 1 to 3.

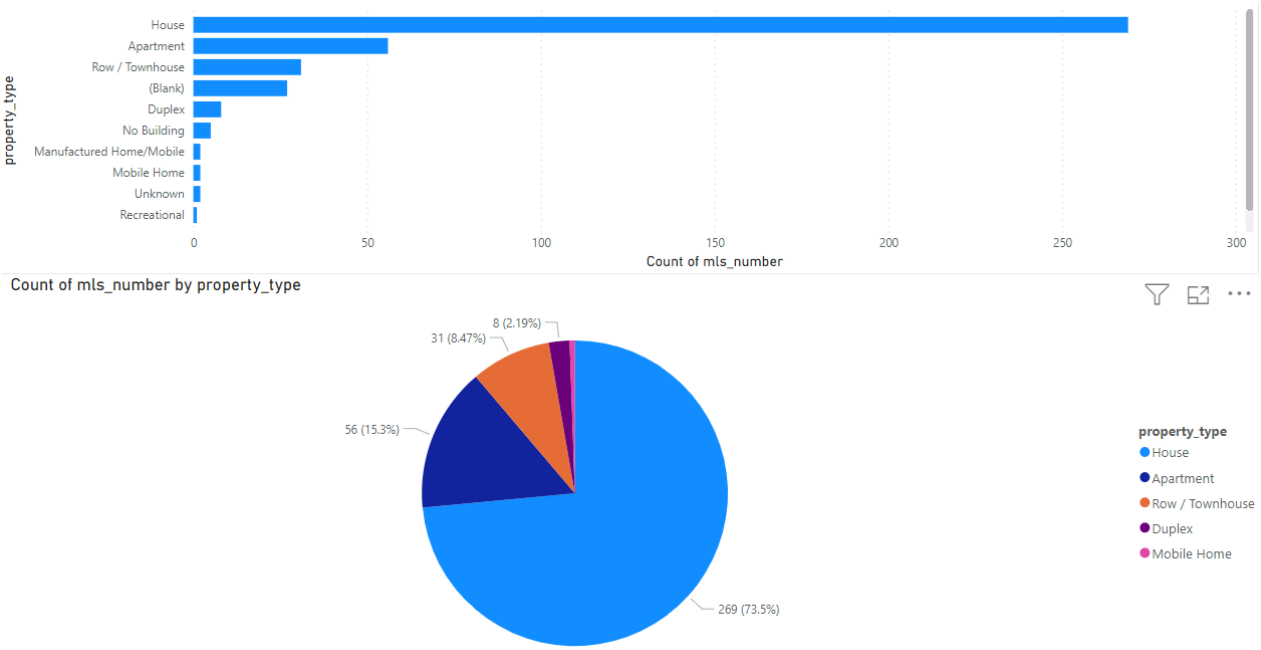


Figure 15. Count of MLS_number by Property_type

Figure 15 shows Sum of MLS_numbers by Property types which are listed on Realtor.ca website. House accounted for 73.50% of Count of mls_number compared to the other Property types.

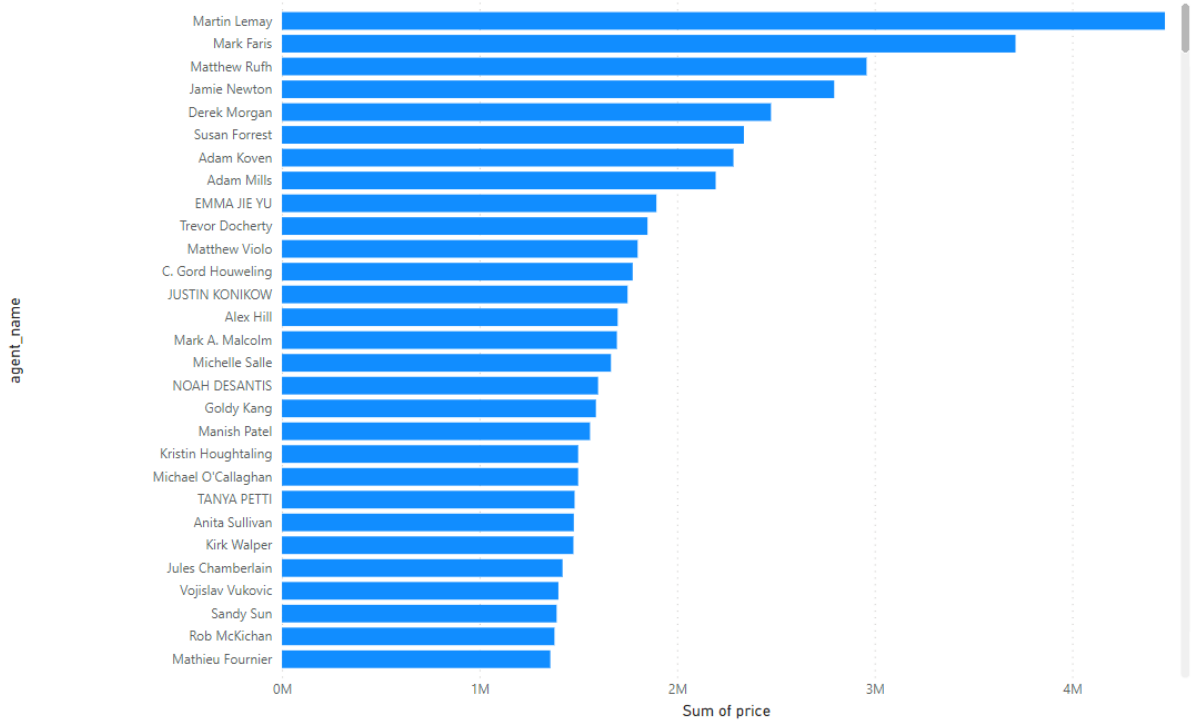
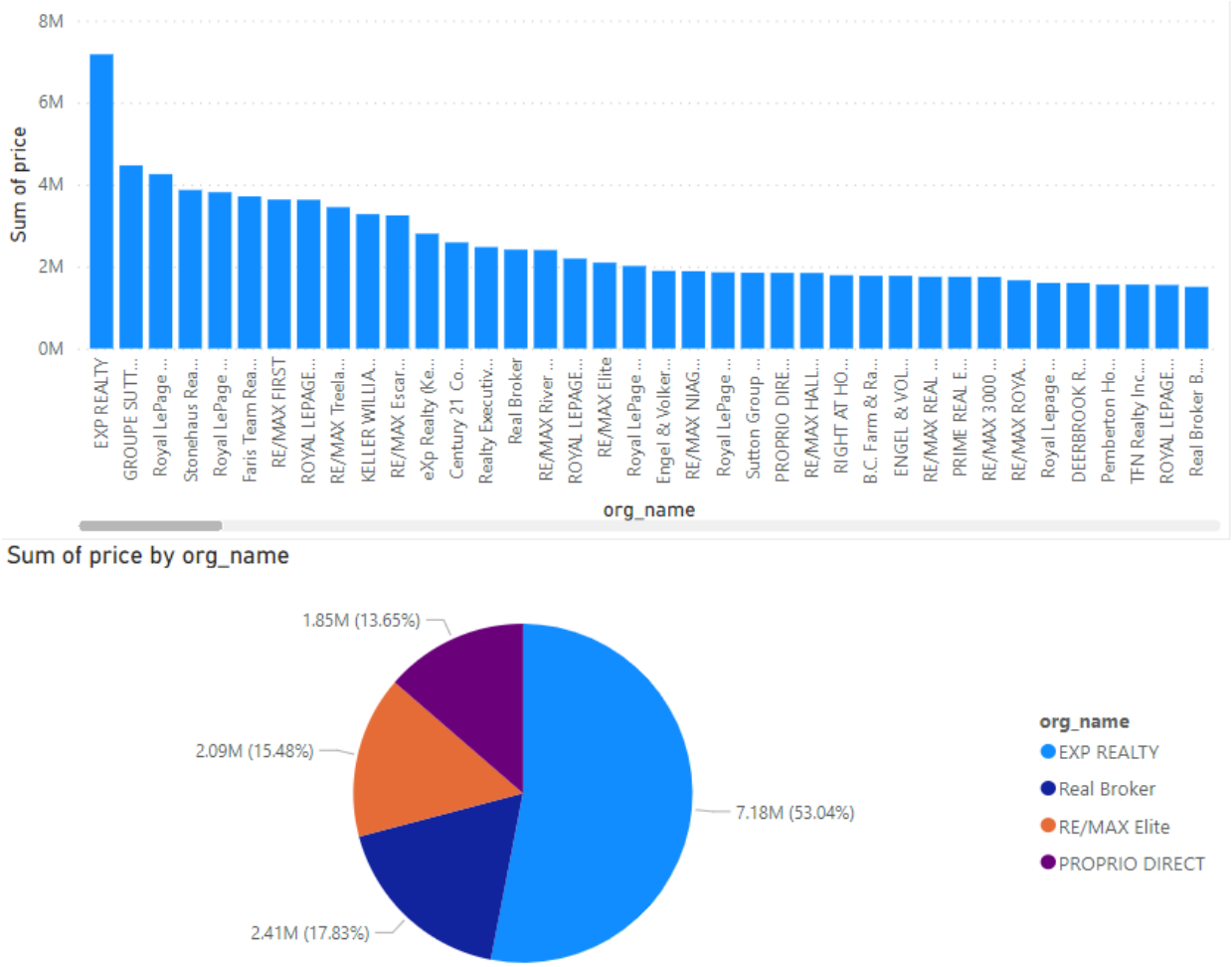


Figure 16. Sum of Price by Agent name

Figure 16 above shows the sum of price by Agent name in which at 4468999, Martin Lemay had the highest Sum of price and was 29,693.33% higher than DWAYNE YOUNG, which had the lowest Sum of price at 15000. Martin Lemay had the highest

Sum of price at 4468999, followed by Mark Faris and Matthew Rufh. DWAYNE YOUNG had the lowest Sum of price at 15000. Martin Lemay accounted for 1.75% of Sum of price. Across all 390 agent_name, Sum of price ranged from 15000 to 4468999.



Sum of price by org_name

Figure 17. Sum of price by Organization name

Figure 17 shows Sum of price by Organization name in which at 7178650, EXP REALTY had the highest Sum of price and was 47,757.67% higher than HomeLife Experts Realty Inc., which had the lowest Sum of price at 15000. EXP REALTY had the highest Sum of price at 7178650, followed by GROUPE SUTTON IMMOBILIA INC. and Royal LePage ProAlliance Realty, Brokerage. HomeLife Experts Realty Inc. had the lowest Sum of price at 15000. EXP REALTY accounted for 2.82% of Sum of price. Across all 298 org_name, Sum of price ranged from 15000 to 7178650.

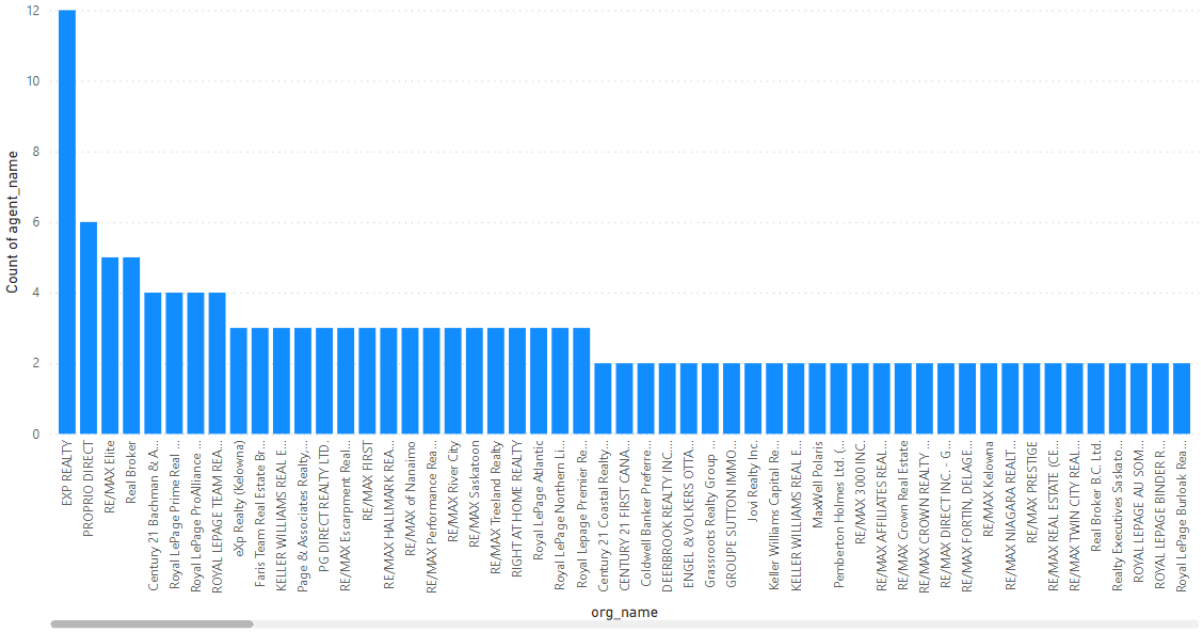


Figure 18. Sum of Agent name by Organization name

Figure 18 above shows the sum of Agent name by Organization name where count of agent_name was highest for EXP REALTY at 12, followed by PROPRIO DIRECT and RE/MAX Elite. EXP REALTY accounted for 2.97% of Count of agent_name. Across all 298 org_name, Count of agent_name ranged from 1 to 12.

5.2 Machine Learning

The evolution of real estate has benefited greatly from machine learning. The results of the model training and testing are satisfying after implementing all the techniques for machine learning described in Chapter 4 section 4.3The algorithms used are listed below, and they are based on the prediction modal described in Chapter 4 above.

5.2.1 Linear Regression

A real estate dataset was used to apply the linear regression model as shown in figure 19, and several criteria were used to assess the model's effectiveness.

```
lin_reg = LinearRegression()
lin_reg.fit(X_train, y_train)
predictions = lin_reg.predict(X_test)

mae, mse, rmse, r_squared = evaluation(y_test, predictions)
print("MAE:", mae)
print("MSE:", mse)
print("RMSE:", rmse)
print("R2 Score:", r_squared)
print("-"*30)
rmse_cross_val = rmse_cv(lin_reg)
print("RMSE Cross-Validation:", rmse_cross_val)

new_row = {"Model": "LinearRegression","MAE": mae, "MSE": mse, "RMSE": rmse, "R2 Score": r_squared, "RMSE (Cross-Validation)": rm
models = models.append(new_row, ignore_index=True)
```

```
MAE: 23567.890565943395
MSE: 1414931404.6297863
RMSE: 37615.57396384889
R2 Score: 0.8155317822983865
------------------------------
RMSE Cross-Validation: 36326.451444669496
```

Figure 19. Linear Regression on Real estate dataset

Above findings shown in figure 19 are acquired as

- **Mean Absolute Error (MAE):** 23,467.89
- **Mean Squared Error (MSE):** 1,414,931,404.63
- **Root Mean Squared Error (RMSE):** 37,615.57
- **R-squared (R2) Score:** 0.8155
- **Cross-Validation Score:** 36,326.45

Overall, the R2 value of 0.8155 suggests that the linear regression model has reasonable prediction ability for the real estate dataset. The MAE, MSE, and RMSE figures, however, indicate that there is potential for improvement in terms of lowering prediction mistakes. The accuracy of real estate price predictions may need to be improved through additional research and model improvement.

5.2.2 Ridge Regression

The provided performance metrics for Ridge Regression applied to a real estate dataset show how effective the model is. Below are the findings which is also shown in figure 20:

- **Mean Absolute Error (MAE):** 23,435.50
- **Mean Squared Error (MSE):** 1,404,264,216.86
- **Root Mean Squared Error (RMSE):** 37,473.51
- **R-squared (R2) Score:** 0.8169
- **Cross-Validation Score:** 35,887.85

```
ridge = Ridge()
ridge.fit(X_train, y_train)
predictions = ridge.predict(X_test)

mae, mse, rmse, r_squared = evaluation(y_test, predictions)
print("MAE:", mae)
print("MSE:", mse)
print("RMSE:", rmse)
print("R2 Score:", r_squared)
print("-"*30)
rmse_cross_val = rmse_cv(ridge)
print("RMSE Cross-Validation:", rmse_cross_val)

new_row = {"Model": "Ridge","MAE": mae, "MSE": mse, "RMSE": rmse, "R2 Score": r_squared, "RMSE (Cross-Validation)": rmse_cross_va
models = models.append(new_row, ignore_index=True)
```

```
MAE: 23435.50371200822
MSE: 1404264216.8595588
RMSE: 37473.513537691644
R2 Score: 0.8169224907874508
------------------------------
RMSE Cross-Validation: 35887.852791598336
```

Figure 20. Ridge Regression on Real estate dataset

The Ridge Regression model seems to perform fairly well on the real estate dataset, achieving a relatively low MAE and RMSE, indicating good predictive accuracy, as well as the R2 score suggesting that a significant portion of the variance in real estate prices is explained by the model, and the cross-validation score further validating the model's performance on unseen data. However, it's important to take into account the specific context of the problem and domain expertise.

5.2.3 Lasso Regression

In a Lasso Regression analysis performed on a real estate dataset, Figure 21 shows the results. Also, the following evaluation metrics were obtained:

- Mean Absolute Error (MAE): 23,560.45
- Mean Squared Error (MSE): 1,414,337,628.50
- Root Mean Squared Error (RMSE): 37,607.68
- R-squared (R2) Score: 0.8156
- Cross-Validation Score: 35,922.77

```
lasso = Lasso()
lasso.fit(X_train, y_train)
predictions = lasso.predict(X_test)

mae, mse, rmse, r_squared = evaluation(y_test, predictions)
print("MAE:", mae)
print("MSE:", mse)
print("RMSE:", rmse)
print("R2 Score:", r_squared)
print("-"*30)
rmse_cross_val = rmse_cv(lasso)
print("RMSE Cross-Validation:", rmse_cross_val)

new_row = {"Model": "Lasso","MAE": mae, "MSE": mse, "RMSE": rmse, "R2 Score": r_squared, "RMSE (Cross-Validation)": rmse_cross_va
models = models.append(new_row, ignore_index=True)
```

```
MAE: 23560.45808027236
MSE: 1414337628.502095
RMSE: 37607.680445649596
R2 Score: 0.815609194407292
------------------------------
RMSE Cross-Validation: 35922.76936876075
```

Figure 21. Lasso regression on Real estate dataset

The low MAE shows that the model's average forecasts are about 23,560.45 off from the actual numbers. The model's forecasts appear to have a spread of roughly 37,607.68, according to the RMSE, a measurement of prediction accuracy.The target variable's variation is explained by the model to some extent, as indicated by the R-squared (R2) score of 0.8156, which indicates a good fit to the data.The model's generalizability is demonstrated by the cross-validation score of 35,922.77, which provides an estimate of how well the model is likely to perform on unseen data.With area for potential development, the Lasso Regression model looks to perform quite well overall on the real estate dataset.

5.2.4 Elastic Net

Elastic Net regression applied to a real estate dataset; the model's performance metrics were as follows:

- **Mean Absolute Error (MAE):** 23,792.74
- **Mean Squared Error (MSE):** 23,792.74
- **Root Mean Squared Error (RMSE):** 41,454.14
- **R-squared (R2) Score:** 0.7759
- **Cross-Validation Score:** 38,449.01

```
elastic_net = ElasticNet()
elastic_net.fit(X_train, y_train)
predictions = elastic_net.predict(X_test)

mae, mse, rmse, r_squared = evaluation(y_test, predictions)
print("MAE:", mae)
print("MSE:", mse)
print("RMSE:", rmse)
print("R2 Score:", r_squared)
print("-"*30)
rmse_cross_val = rmse_cv(elastic_net)
print("RMSE Cross-Validation:", rmse_cross_val)

new_row = {"Model": "ElasticNet","MAE": mae, "MSE": mse, "RMSE": rmse, "R2 Score": r_squared, "RMSE (Cross-Validation)": rmse_cro
models = models.append(new_row, ignore_index=True)
```

```
MAE: 23792.743784996732
MSE: 1718445790.1371393
RMSE: 41454.14080809225
R2 Score: 0.775961837382229
------------------------------
RMSE Cross-Validation: 38449.00864609558
```

Figure 22. Elastic Net on Real Estate dataset

Overall, these metrics imply that the Elastic Net regression model has been applied to the real estate dataset, and although it provides a good fit and explains a sizable portion of the variance, there is room for improvement in terms of lowering prediction errors, especially for high-value properties or outliers in the dataset. The model's performance might be improved with additional optimisation and feature engineering.

5.2.5 Support Vector Machines

Support Vector Machines (SVM) applied to a real estate dataset achieved promising results which is shown in figure 23. Below are the results:

- **Mean Absolute Error (MAE):** 17,843.16
- **Mean Squared Error (MSE):** 1,132,136,370.34
- **Root Mean Squared Error (RMSE):** 33,647.23
- **R-squared (R2) Score:** 0.8524
- **Cross-Validation Score:** 30,745.47

```
svr = SVR(C=100000)
svr.fit(X_train, y_train)
predictions = svr.predict(X_test)

mae, mse, rmse, r_squared = evaluation(y_test, predictions)
print("MAE:", mae)
print("MSE:", mse)
print("RMSE:", rmse)
print("R2 Score:", r_squared)
print("-"*30)
rmse_cross_val = rmse_cv(svr)
print("RMSE Cross-Validation:", rmse_cross_val)

new_row = {"Model": "SVR","MAE": mae, "MSE": mse, "RMSE": rmse, "R2 Score": r_squared, "RMSE (Cross-Validation)": rmse_cross_val}
models = models.append(new_row, ignore_index=True)
```

```
MAE: 17843.16228084976
MSE: 1132136370.3413317
RMSE: 33647.234215330864
R2 Score: 0.852400492526574
------------------------------
RMSE Cross-Validation: 30745.475239075837
```

Figure 23. Support Vector Machines on Real estate dataset

These findings suggest that the SVM model has high prediction ability on the real estate dataset, captures a sizeable amount of the price variance, and provides insightful information for real estate-related decision-making. However, it's crucial to

evaluate these findings in light of the particular objectives and conditions of the real estate application.

5.2.6 Random Forest Regression

In a Random Forest Regression analysis conducted on a Real Estate dataset, Figure 24 shows the results of Random Forest Regressor on the Real estate dataset. Also, the following evaluation metrics were obtained:

- **Mean Absolute Error (MAE):** 17,843.16
- **Mean Squared Error (MSE):** 1,132,136,370.34
- **Root Mean Squared Error (RMSE):** 33,647.23
- **R-squared (R2) Score:** 0.8524
- **Cross-Validation Score:** 30,745.47

```
random_forest = RandomForestRegressor(n_estimators=100)
random_forest.fit(X_train, y_train)
predictions = random_forest.predict(X_test)

mae, mse, rmse, r_squared = evaluation(y_test, predictions)
print("MAE:", mae)
print("MSE:", mse)
print("RMSE:", rmse)
print("R2 Score:", r_squared)
print("-"*30)
rmse_cross_val = rmse_cv(random_forest)
print("RMSE Cross-Validation:", rmse_cross_val)

new_row = {"Model": "RandomForestRegressor","MAE": mae, "MSE": mse, "RMSE": rmse, "R2 Score": r_squared, "RMSE (Cross-Validation)
models = models.append(new_row, ignore_index=True)
```

```
MAE: 18115.11067351598
MSE: 1004422414.0219476
RMSE: 31692.623968708358
R2 Score: 0.869050886899595
------------------------------
RMSE Cross-Validation: 31138.863315259332
```

Figure 24. Random Forest Regressor on Real Estate Dataset

An overall low MAE and RMSE, a good R2 score, and a generally consistent cross-validation score suggest that the Random Forest Regression model is doing well on the Real Estate dataset. These findings imply that the model is highly accurate in predicting real estate values, accounting for a sizeable amount of the price fluctuation.

5.2.7 XG Boost Regressor

The XG Boost regressor was used in this investigation to apply to a real estate dataset as shown in figure 25, and the model's performance indicators were assessed:

- **Mean Absolute Error (MAE):** 17,439.91
- **Mean Squared Error (MSE):** 716,579,004.52
- **Root Mean Squared Error (RMSE):** 26,768.99
- **R-squared (R2) Score:** 0.9065
- **Cross-Validation Score:** 29,698.84

```
xgb = XGBRegressor(n_estimators=1000, learning_rate=0.01)
xgb.fit(X_train, y_train)
predictions = xgb.predict(X_test)

mae, mse, rmse, r_squared = evaluation(y_test, predictions)
print("MAE:", mae)
print("MSE:", mse)
print("RMSE:", rmse)
print("R2 Score:", r_squared)
print("-"*30)
rmse_cross_val = rmse_cv(xgb)
print("RMSE Cross-Validation:", rmse_cross_val)

new_row = {"Model": "XGBRegressor","MAE": mae, "MSE": mse, "RMSE": rmse, "R2 Score": r_squared, "RMSE (Cross-Validation)": rmse_c
models = models.append(new_row, ignore_index=True)
```

```
MAE: 17439.918396832192
MSE: 716579004.5214689
RMSE: 26768.993341578403
R2 Score: 0.9065777666861116
------------------------------
RMSE Cross-Validation: 29698.84961808251
```

Figure 25. XG Boost Regressor on Real Estate Dataset

In conclusion, the XG Boost regressor appears to perform well on the real estate dataset, with a relatively low MAE and RMSE, a high R2 score, and a reasonable cross-validation score. These results suggest that the model is effective at predicting real estate prices, capturing a significant portion of the underlying variance in the data. However, further analysis and domain-specific considerations may be necessary to determine the practical utility of the model in real-world applications.

5.2.8 Polynomial Regression

In a polynomial regression analysis conducted on a real estate dataset as shown In figure 26, the model's performance metrics indicate significant issues:

- **Mean Absolute Error (MAE):** 17,439.91
- **Mean Squared Error (MSE):** 716,579,004.52
- **Root Mean Squared Error (RMSE):** 26,768.99
- **R-squared (R2) Score:** 0.9065
- **Cross-Validation Score:** 29,698.84

```
poly_reg = PolynomialFeatures(degree=2)
X_train_2d = poly_reg.fit_transform(X_train)
X_test_2d = poly_reg.transform(X_test)

lin_reg = LinearRegression()
lin_reg.fit(X_train_2d, y_train)
predictions = lin_reg.predict(X_test_2d)

mae, mse, rmse, r_squared = evaluation(y_test, predictions)
print("MAE:", mae)
print("MSE:", mse)
print("RMSE:", rmse)
print("R2 Score:", r_squared)
print("-"*30)
rmse_cross_val = rmse_cv(lin_reg)
print("RMSE Cross-Validation:", rmse_cross_val)

new_row = {"Model": "Polynomial Regression (degree=2)","MAE": mae, "MSE": mse, "RMSE": rmse, "R2 Score": r_squared, "RMSE (Cross-
models = models.append(new_row, ignore_index=True)
```

```
MAE: 2382228327828308.5
MSE: 1.5139911544182342e+32
RMSE: 1.230443478758059e+16
R2 Score: -1.9738289005226644e+22
------------------------------
RMSE Cross-Validation: 36326.451444669496
```

Figure 26. Polynomial Regression on Real Estate Dataset

In conclusion, the findings of this polynomial regression study on the real estate dataset are incredibly peculiar and point to serious flaws in the model. It is likely that

there are problems with the data, the model's specification, or both because the model's predictions are far from accurate. The effectiveness of the model needs to be increased, maybe by additional research or an alternative modelling strategy.

5.3 Model Comparison

A critical phase in the model construction process in machine learning is model comparison. It entails comparing various machine learning algorithms or iterations of the same algorithm to establish which one excels at a certain task. Depending on the nature of the issue (classification, regression, clustering, etc.), this comparison is frequently conducted using a variety of evaluation metrics, including accuracy, precision, recall, F1-score, and more.

Choosing the model with the best predicted performance and good generalisation to new data is the aim of model comparison. Making informed choices regarding which model to use in practical applications aids data scientists and machine learning professionals. Additionally, model comparison assists in spotting possible over- or underfitting problems and fine-tuning hyperparameters for optimum performance.

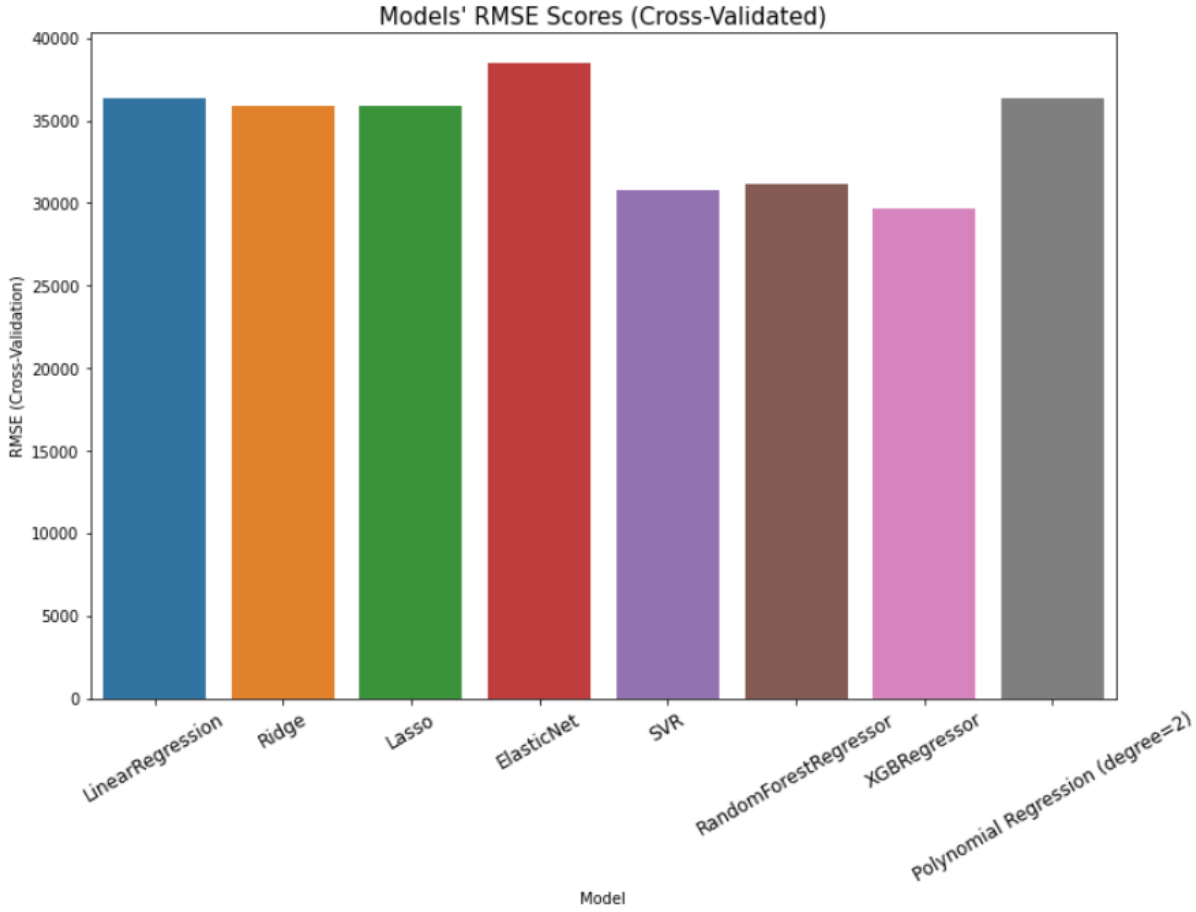


Figure 27. Model comparison of all the algorithms used in this Real Estate Project

The result from the comparison clearly shows that Elastic Net has the higher cross validation then the other algorithms applied.

Chapter 6: Conclusions and Future Work

6.1 Conclusions

This Research aims to investigate the use of business intelligence and machine learning in the Canadian real estate sector. looking at the many tactics, methods, and tools employed in this scenario and highlight both their benefits and drawbacks. The data from the ETL generates an interesting report in PowerBI. Various PowerBI reports explain the significance in the real estate sector. The geolocation in this study displays every location in Canada where real estate is for sale when it comes to a visualisation of the properties. The majority of the properties listed on realtor.ca are 3-bedroom property and the property with 2 bathrooms. The number of agents who have posted advertisements on the website; Mark Faris agent is having more advertisements; and the types of properties listed on the website; houses are more frequently listed than the other types of properties on the website. Martin Lemay had the highest Sum of price and was 29,693.33% higher than DWAYNE YOUNG, which had the lowest Sum of price at 15000. The Organisation EXP REALTY had the highest Sum of price and was 47,757.67% higher than HomeLife Experts Realty Inc. Also, the count of agent_name was highest for EXP REALTY at 12 as compared to others.

Regarding machine learning, several methods were used with remarkably satisfying results on the real estate dataset. We concluded that Elastic.Net has greater cross validation compared to the other algorithms used after applying machine learning techniques and model building. By project's end, a satisfying comparison had been made as 2 of the algorithms i-e Elastic Net and Polynomial Regression are higher than the other.

6.2 Future Work

Even after all the effort and thought put into the conclusion, a more accurate algorithm than the ones employed in this thesis may yet be developed, or a new algorithm that provides insightful real estate data may be discovered. We can automate this entire process in light of the fact that the data taken was only current data, schedule the automation for daily execution, and get updated insights from the advertising posted on Realtor.ca. We can also store the data for future analytics in database.
Also, we can make our own website and link it with the database of this project upon which customers can easily search for house of their need. A website made on PHP can be linked to this database by simply putting trigger command on the relevant tables.

References

Agarwal, S. (2022). Mortgage Refinancing, Consumer Spending, and Competition: Evidence from the Home Affordable Refinance Program.

Andrew Baum, A. S. (2020). The future of real estate.

Aziz, N. &. (2019). Predicting Housing Prices: A Comparison of Four Models. 245-272.

Bao, Y. H. (2018). Predicting House Prices with Machine Learning Models. 1043-1050.

Campoli, F. (n.d.). *PostgreSQL for DBA: PostgreSQL 12.* 2020.

Canada, G. o. (n.d.). Retrieved from Government of Canada: https://www.canada.ca/en.html

Chen, L. &. (2019). Data Integration for Real Estate Management. *International Conference on Artificial Intelligence and Industrial Automation.* Chen, L., & Zhang, C.

Dumitru-Alexandru, B. (2016). Business Intelligence for decision making in Economics. 125--158.

estate, C. M. (2022). Retrieved from Canada Mortgage and Housing Corporation in real estate: https://www.cmhc-schl.gc.ca

Kakkar, M. &. (2019). Machine Learning in Real Estate: A Comprehensive Review. 225-229.

Li, J. W. (2020). Real Estate Data Warehouse Design and Implementation. *International Conference on Data Science and Advanced Analytics.* Li, J., Wang, Y., & Zhao, Y.

Ma, H. Z. (2018). Challenges and Opportunities of Data Integration in Real Estate Transactions. *IEEE International Conference on Big Data.* Ma, H., Zhang, P., & Li, Q.

Na Li, R. Y. (2020). Factors Affect the Housing Prices in China.

Roldán, M. C. (2013). *Pentaho Data Integration Beginner's Guide.*

Saini, C. a. (2016). Information retrieval in web crawling: A survey. In C. a. Saini, *Proceedings of the 2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI'16).* IEEE.

SALAM FRAIHAT, W. A. (2021). Business Intelligence Framework Design and Implementation: A Real-estate Market Case Study. 1–16.

Scholly, É. (2019). Business Intelligence & analytics applied to public housing. 552--557.

Şeker, S. E. (2018). Predicting Real Estate Prices with Multiple Regression Analysis. *International Conference on Advanced Computer Science and Information Systems*, (pp. 1-6).

Sherman, R. (2014). *Business Intelligence Guidebook.* Morgan Kaufmann.

Smith, J. (2019). Leveraging Business Intelligence in the Canadian Real Estate Market. 43--52.

Smith, J. A. (2020). The Role of Business Intelligence in the Real Estate Sector: A Comprehensive Review. 155-176.

Tanwar, S. &. (2017). Business Intelligence and Real Estate: A Comprehensive Review. 6-11.

Thompson, A. C. (2020). Machine Learning Applications in Canadian Real Estate: A Review. International Journal of Real Estate Studies. 75--93.